Cognitive Psychology 110 (2019) 30–69

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

How the input shapes the acquisition of verb morphology: Elicited production and computational modelling in two highly inflected languages

Felix Engelmann[a,e,*], Sonia Granlund[a,b,e], Joanna Kolak[a,c,e], Marta Szreder[a,d,e], Ben Ambridge[b,e], Julian Pine[b,e], Anna Theakston[a,e], Elena Lieven[a,e]

[a] University of Manchester, UK
[b] University of Liverpool, UK
[c] University of Warsaw, Poland
[d] United Arab Emirates University, United Arab Emirates
[e] ESRC International Centre for Language and Communicative Development (LuCiD), UK

ARTICLE INFO

Keywords:
Language acquisition
Verb marking
Morphology
Computational modelling
Cross-linguistic
Elicited production
Neural networks

ABSTRACT

The aim of the present work was to develop a computational model of how children acquire inflectional morphology for marking person and number; one of the central challenges in language development. First, in order to establish which putative learning phenomena are sufficiently robust to constitute a target for modelling, we ran large-scale elicited production studies with native learners of Finnish (N = 77; 35–63 months) and Polish (N = 81; 35–59 months), using a novel method that, unlike previous studies, allows for elicitation of all six person/number forms in the paradigm (first, second and third person; singular and plural). We then proceeded to build and test a connectionist model of the acquisition of person/number marking which not only acquires near adult-like mastery of the system (including generalisation to unseen items), but also yields all of the key phenomena observed in the elicited-production studies; specifically, effects of token frequency and phonological neighbourhood density of the target form, and a pattern whereby errors generally reflect the replacement of low frequency targets by higher-frequency forms of the same verb, or forms with the same person/number as the target, but with a suffix from an inappropriate conjugation class. The findings demonstrate that acquisition of even highly complex systems of inflectional morphology can be accounted for by a theoretical model that assumes rote storage and phonological analogy, as opposed to formal symbolic rules.

1. Introduction

One of the most complex challenges that children (and indeed adults) commonly face when acquiring a language is mastering systems of inflectional morphology; systems that — amongst other things — mark verbs for person (e.g., *I like* [1st person] vs. *He likes* [3rd person]) and number (e.g., *He likes* [singular] vs. *They like* [plural]). Inflectional morphology occupies a central place in language acquisition research for two reasons. First, systems are often extremely complex — it is not uncommon for languages to have tens or even hundreds of different person/number markers (e.g., Stoll et al., 2012) — meaning that their acquisition is one of the most crucial challenges facing learners. Second — perhaps in part because of their complexity — systems of inflectional morphology have

* Corresponding author at: Child Study Centre, Coupland 1, University of Manchester, Oxford Road, Manchester M13 9PL, UK.
E-mail address: felix.engelmann@manchester.ac.uk (F. Engelmann).

https://doi.org/10.1016/j.cogpsych.2019.02.001
Accepted 3 February 2019

traditionally constituted a useful test case for different accounts of language acquisition more generally; in particular, accounts based on formal context-free symbolic rules (e.g., Pinker, 1998; Pinker & Ullman, 2002) and those based on storage of, and analogy across, individual forms (e.g., Bybee & Moder, 1983; McClelland & Patterson, 2002; Tomasello, 2009).

Within the domain of inflectional morphology, a particularly key question is how learners acquire systems used to mark verbs for both person (1st = speaker, 2nd = listener, 3rd = neither speaker nor listener) and number (one speaker/listener/third party, or several). Verb person/number marking is of central importance to theories of morphological acquisition and of language acquisition more generally, because these systems tend not only to be relatively complex, but also relatively common typologically: The World Atlas of Linguistic Structures lists around 300 languages (including almost all Indo-European languages) that make use of such a system (Dryer & Haspelmath, 2013).

It is therefore surprising that, although many verbal and computational accounts of morphological acquisition have been proposed (these are set out in detail below), the vast majority have focused on considerably simpler — and typologically-speaking, much rarer — systems, such as English past-tense marking. Indeed, to our knowledge, there currently exists no computational model that attempts to simulate the acquisition of person/number marking in a highly inflected language. Our goal in the present work is therefore to investigate whether theoretical and modelling approaches developed for these simpler systems can scale-up to yield a computational model of the more complex — yet more typologically common and theoretically crucial — phenomenon of verb person/number marking.

Computational models of morphological acquisition are generally evaluated against three criteria. The first is that the model must eventually — like human learners — achieve essentially perfect mastery of the system. Although native-speaking adults do occasionally produce verb agreement errors (e.g., Bock & Miller, 1991), they generally do so only in cases of long-distance dependencies (e.g., ∗*The key to the cabinets are missing*), and not in the types of simple sentences with which we are concerned in the present study. Second, and relatedly, this mastery must include the ability to produce correctly inflected forms that have not been witnessed in the input. Language dictionaries typically list somewhere in the region of 10,000 – 20,000 verbs (Oxford English Dictionary, 2018), the majority of which are of extremely low frequency, and will have been witnessed in just one or two of their possible person/number forms. Yet adults, and even older children, are able to inflect low frequency or even novel forms with essentially perfect accuracy (e.g., Berko, 1958; Savičiūtė, Ambridge, & Pine, 2018). Third, and finally, a successful model of morphological acquisition must be able to simulate not only the adult endpoint, but also all well-established phenomena relating to patterns of correct use and error shown by children.

1.1. Phenomena observed in previous studies of inflectional morphology

1.1.1. Overall error rates are low, but are high for rare person/number contexts

The earliest investigations of children's acquisition of verbal person/number morphology generally involved analysis of children's spontaneous speech data (Deen, 2004; Harris & Wexler, 1996; Hoekstra & Hyams, 1998; Smoczyńska, 1985; Wexler, 1998; although see MacWhinney, 1978; Slobin & Bever, 1982 for early experimental studies). These studies — regardless of theoretical perspective — generally reported that, while omissions of person/number marking (e.g., ∗*He like cake* for *He likes cake*) are common in many languages, errors of commission (i.e., using an incorrect form of the verb) are relatively rare. For example, Hoekstra and Hyams (1998) found that across Spanish, Italian, German and Catalan, such errors were produced at rates below 5%.

Subsequently, more detailed naturalistic data analyses (e.g., Aguado-Orea, 2004; Aguado-Orea & Pine, 2015; Rubino & Pine, 1998) found that, although overall error rates are low, these predominantly reflect performance with certain person/number marked forms of particular verbs that children use very often, and know very well (e.g., Spanish *quiero*, 'I want'). Person/number contexts that children hear and use rarely (e.g. 3pl) often have error rates as high as 50%, but, because these forms are used so rarely, this barely affects overall error rates that collapse across person/number contexts.

Broadly similar findings have been reported in elicited production studies conducted in Italian (Leonard, Caselli, & Devescovi, 2002), Hungarian (Leonard, Kas, & Pléh, 2009), Greek (Stavrakaki & Clahsen, 2009) and Finnish (Kirjavainen, Nikolaev, & Kidd, 2012; Kunnari et al., 2011; Räsänen, Ambridge, & Pine, 2016). Again, overall error rates that are relatively low (if generally somewhat higher than those found in naturalistic data, due to broader coverage of the system), mask much higher error rates in lower frequency person/number contexts. For example, Räsänen et al. (2016) reported an overall error rate of around 15%, but an error rate of 36% for 2pl verb forms, as compared with just 0.5% for 1pl verb forms. If, as we anticipate, this pattern is confirmed in the present empirical studies, it is an important phenomenon to simulate in our model.

1.1.2. Errors are more common for individual inflected forms with low token frequency

Surprisingly few previous studies have investigated error rates at the level of individual inflected verb forms (e.g., Finnish *nukun*, 'I sleep' [1sg] vs. *nukut*, 'You sleep' [2sg]), as opposed to at the level of person/number contexts collapsing across the identity of the verb (see Section 1.1.1). However, wherever this has been done (often in the domain of English past-tense marking), researchers have found a negative correlation between token frequency in the input and the rate of error in children's production, whether analysing naturalistic (e.g., Aguado-Orea, 2004; Aguado-Orea & Pine, 2015; Maslen, Theakston, Lieven, & Tomasello, 2004) or experimental elicited production data (e.g., Ambridge, 2010; Leonard et al., 2002; Marchman, 1997; Matthews & Theakston, 2006; Räsänen et al., 2016; Tatsumi, Ambridge, & Pine, 2017; Theakston, Lieven, & Tomasello, 2003; Theakston & Rowland, 2009), including in studies of noun case-marking in Polish (Dąbrowska, 2008a; Dąbrowska & Szczerbiński, 2006), Serbian (Mirković, Seidenberg, & Joanisse, 2011), and Lithuanian (Savičiūtė et al., 2018). This finding is unsurprising, given that effects of token frequency are ubiquitous across many, perhaps all, domains of language acquisition (Ambridge, Kidd, Rowland, & Theakston, 2015), and we therefore expect to replicate it in the present elicitation studies.

1.1.3. Many errors (usually a clear majority) are of one of three types: (a) frequency-based substitutions; (b) near-miss errors; (c) conjugation-class errors

Again, surprisingly few studies have examined in detail exactly what children do when they fail to correctly produce a particular person/number target form. Those that have done so have found that (a) a large proportion of errors of commission (e.g., 42% in the study of Räsänen et al., 2016) involve the replacement of low-frequency target forms with a higher-frequency form of the same verb (e.g., Aguado-Orea, 2004; Aguado-Orea & Pine, 2015; Leonard et al., 2002; Marchman, 1997; Matthews & Theakston, 2006; Räsänen et al., 2016; Rubino & Pine, 1998; Theakston et al., 2003; Theakston & Rowland, 2009; see also Dąbrowska, 2008a; Dąbrowska & Szczerbiński, 2006; Dąbrowska & Tomasello, 2008; Savičiūtė et al., 2018, for noun case marking). However, by no means all errors are of this type. Also common are (b) near-miss errors (Leonard et al., 2002) where children produce a form bearing the correct person but incorrect number marking (e.g., 3sg in place of 3pl) or vice versa (3sg in place of 2sg); though — as these examples show — such errors are often hard to distinguish from frequency-based substitutions. Finally, (c) conjugation class errors occur when children use an inflectional ending that exhibits the same person and number as the target form, but is from a different conjugation class (e.g. in Polish using *umią* [incorrectly applied ending from 3rd plural, class II] in place of *umieją*, 'they can' [3rd plural, class IV]). These errors, sometimes referred to as examples of "inflectional imperialism" (Slobin, 1968), have mainly been observed in studies of noun case marking (e.g., Gvozdev, 1949, for Russian, Savičiūtė et al., 2018, for Lithuanian, Dąbrowska, 2004, 2005, 2008a; Krajewski, Lieven, & Theakston, 2012, for Polish), but are also expected to occur for verb person/number inflections, particularly for languages such as Polish which have a large number of conjugation classes.

Given that this error pattern — whereby a clear majority of errors usually constitute one of these three types — has been observed across several studies in different domains, we again anticipate that this finding will replicate in the present empirical studies, and so constitute an important phenomenon for subsequent simulation.

1.1.4. Errors are less common for verbs that score high on phonological neighbourhood density (PND)/type frequency/islands of reliability/class size

A common claim in the literature is that children show higher rates of correct production — and hence lower error rates — for forms with many phonological neighbours (as discussed in the Methods section, this can be defined in a number of different ways). For example, in the domain of English past-tense inflection, over-regularization errors (e.g., ∗*sleeped*) are rare for verbs with many phonological neighbours: verbs that have stems with similar-sounding endings, and that form the past-tense in the same way (e.g., *keep/kept*; *creep/crept*; *weep/wept*). Intuitively, the idea is that, if the correct target cannot be retrieved from memory, a large phonological neighbourhood increases the probability of finding a suitable form for analogy (e.g., if *keep* →*kept* then *sleep* →*slept*). Although there is considerable evidence for this claim in the domain of English past-tense marking (e.g., Albright & Hayes, 2003; Ambridge, 2010; Blything, Ambridge, & Lieven, 2018; Marchman, 1997; Marchman, Wulfeck, & Weismer, 1999; Matthews & Theakston, 2006; Prasada & Pinker, 1993) and noun case marking in Polish (Dąbrowska, 2008a), Serbian (Mirković et al., 2011) and Lithuanian (Savičiūtė et al., 2018), few studies have investigated this phenomenon in the domain of person/number marking (Räsänen et al., 2016, see also Finnish past tense Kirjavainen et al., 2012). Nevertheless, because this effect has been frequently observed in related domains, this is again an effect that we expect to replicate here, and therefore to constitute a phenomenon for subsequent modelling (for further effects of both frequency and class size in Finnish in a very different experimental context, see Bertram, Laine, & Virkkala, 2000).

1.1.5. Phonological neighbourhood density (PND) effects may be smaller for forms with higher token frequency

Räsänen et al. (2016) observed a negative interaction between token frequency (Section 1.1.2) and phonological neighbourhood density (PND) (Section 1.1.4), such that a larger PND had a greater facilitative effect for lower frequency than higher frequency target verb forms. The authors argued that this interaction reflects the fact that high-frequency target forms can easily be retrieved by rote, obviating the need for phonological analogy. It is only lower-frequency target forms, which are not easily retrieved, that require generation by phonological analogy, and so benefit from membership of a larger phonological neighbourhood. This explanation is plausible, not least because this kind of *frequency* × *regularity* interaction has also been found in other domains of language processing. For example, in word recognition and naming, effects of phonological sub-regularities for irregular items are observed primarily for low-frequency words (Andrews, 1982; Seidenberg, 1985; Seidenberg, Waters, Barnes, & Tanenhaus, 1984; Taraban & McClelland, 1987; Waters & Seidenberg, 1985). Similar interactions have been found for frequency and regularity of sentence structures in comprehension (MacDonald & Christiansen, 2002; Pearlmutter & MacDonald, 1995). Despite this, we are not aware of any other study of inflectional morphology except Räsänen et al. (2016) that has observed such an effect. Furthermore, even in Räsänen et al. (2016), this interaction was observed only in a statistical model that did not include age and its interactions, and the token-frequency counts were taken from newspaper corpora (as opposed to child-directed speech). It is therefore important for the present empirical studies to establish more definitively whether or not this interaction reflects a genuine phenomenon that must be captured by our model.

1.1.6. Effects of phonological neighbourhood density and token frequency may decrease with age

Räsänen et al. (2016) also observed a negative interaction between age and PND, reflecting a decrease in the importance of PND with increasing age. The authors argued that this finding is due to learners' knowledge of the system becoming increasingly abstract with development, leaving them less reliant on analogy with close phonological neighbours. However, this explanation is not entirely convincing, given that the interaction has not been observed in any other studies, and moreover, several of the studies listed in Section 1.1.4 (e.g., Albright & Hayes, 2003; Ambridge, 2010; Blything et al., 2018; Prasada & Pinker, 1993) — albeit studies of the

Table 1
Examples of person/number marking in verbs from each major verb class in Finnish, along with its 1st infinitive form and English translation.

Class English	I 'to sleep'	II 'to get'	III 'to come'	IV 'to jump'
1st inf	nukkua	saada	tulla	hypätä
1sg	nuku-n	saa-n	tule-n	hyppää-n
2sg	nuku-t	saa-t	tule-t	hyppää-t
3sg	nukku-u	saa-0	tule-e	hyppää-0
1pl	nuku-mme	saa-mme	tule-mme	hyppää-mme
2pl	nuku-tte	saa-tte	tule-tte	hyppää-tte
3pl	nukku-vat	saa-vat	tule-vat	hyppää-vät

English past-tense — observed PND effects for children as old as 9–10, and even adults.

Räsänen et al. (2016) did not observe an interaction between age and token frequency. On the one hand, such an interaction might be expected, on the basis that older children have more abstract knowledge, and so rely less on frequency-sensitive retrieval of stored forms. On the other hand, many studies (see Ambridge et al., 2015 for a review) have observed lexical frequency effects in adults, meaning that one would not necessarily expect to see such an interaction.

In summary, previous studies offer little to no evidence regarding whether effects of PND and token frequency decrease with age; we therefore leave these as open questions to be answered by our empirical elicitation studies.

1.2. Previous theoretical and computational approaches

Early theoretical accounts of morphological acquisition (e.g., Hoekstra & Hyams, 1998; Pinker, 1984; Wexler, 1998) generally adopted a generativist-nativist framework, under which the learner's task is to fill in the cells of an empty, innately-specified, person/number marking paradigm (e.g., Table 1) with the relevant inflectional morphemes (e.g., Finnish 1sg = *n*; 2sg = *t*, 3sg = *u*), for example by using a hypothesis-testing, affix-stripping, procedure (Pinker, 1984). Forms are then produced by the application of a formal rule that checks off the person and number features of the verb form, but is insensitive to its lexical content (though as we will see shortly, more recent generativist-nativist accounts have increasingly posited roles for rote storage/retrieval and phonological analogy).

In contrast, connectionist models aim to produce rule-like behaviour using a mechanism that "does not contain any statement of this rule" (Rumelhart & McClelland, 1985: 217). Rumelhart and McClelland (1985) offer the example of a honeycomb, which could be described using a mathematical rule, but in fact "arises from the interaction of forces that wax balls exert on each other when compressed" (p. 217). Similarly, modern day "deep-learning" networks, successors to this early parallel distributed processing approach (Rumelhart, McClelland, & the PDP Research Group, 1986) achieve good performance in visual object recognition, not by using formal rules, but "by the simultaneous processing of multiple individually inconclusive cues" (Rogers & McClelland, 2014: 1027). Through gradual error correction by repeated exposure to examples, the network weights become adjusted in a way that makes optimal use of the available cues. Rule-like behaviour emerges in contexts where the error is affected by only a small number of cues. An important property of PDP models is that they generally show the kind of frequency × regularity interaction discussed in Section 1.1.5 (*Phonological neighbourhood density (PND) effects may be smaller for forms with higher token frequency*). Most prominently, this has been demonstrated with models of word recognition (Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989) and sentence processing (Juliano & Tanenhaus, 1994; MacDonald & Christiansen, 2002).

Beginning with Rumelhart and McClelland (1985), many different models have applied this approach to morphological acquisition (e.g., Cottrell & Plunkett, 1994; Daugherty & Seidenberg, 1992, 1994; Elman, 1998; Forrester & Plunkett, 1994; Hahn & Nakisa, 2000; Hare & Elman, 1995; Hare, Elman, & Daugherty, 1995; Joanisse & McClelland, 2015; Joanisse & Seidenberg, 1999; MacWhinney & Leinbach, 1991; Mirković et al., 2011; Plunkett & Bandelow, 2006; Plunkett & Juola, 1999; Plunkett & Marchman, 1991, 1993, 1996; Plunkett & Nakisa, 1997; Plunkett, Sinha, Møller, & Strandsby, 1992; Ruh & Westermann, 2008; Rumelhart & McClelland, 1985; Seidenberg & McClelland, 1989; Westermann & Ruh, 2012). These models are closely related to computational and linguistic exemplar models (e.g., Bybee, 1985, 1995; Bybee & Moder, 1983; Chandler, 2010), which also largely eschew formal symbolic rules, with learners instead relying on rote storage/retrieval of phonological forms (paired with their meanings), and phonological analogy across them.

A question that remains controversial is whether such models can succeed in acquiring systems of inflectional morphology while eschewing formal symbolic rules in their entirety (e.g., Marcus, 2001). For example, Rumelhart and McClelland's (1985) model learned probabilistic associations between patterns that represented the phonological properties of English verb stems and their past-tense forms, which allowed the model to discover partial phonological regularities below the level of a rule. The model exhibited typical developmental error patterns, such as the U-shaped learning trajectory, with an increasing rate of overgeneralisations that are eventually overcome. However, critics of the approach pointed out that Rumelhart and McClelleland's model did not in fact achieve adultlike performance (Pinker & Prince, 1988), and that many subsequent attempts to address these shortcomings (e.g., MacWhinney & Leinbach, 1991; Plunkett & Marchman, 1991, 1993, 1996) "build in or presuppose surrogates for the linguistic phenomena they claim to eschew" (Pinker & Ullman, 2002), such as the notion of a verb stem (e.g., *play*) and regular morpholgical marking (e.g., *-ed*). On the other hand, while models based on formal symbolic rules generally offer better coverage of the data (e.g., O'Donnell, 2015;

Yang, 2002), they typically achieve this coverage at the expense of explaining how the relevant generalisations are actually acquired. Both of these models were supplied with regular and irregular rules for transforming English verb stems into past-tense forms; their task was simply to learn which rule applies to each verb. Given that the debate over the necessity of formal symbolic rules remains unresolved, the extent to which the present model incorporates such rules — even if only indirectly in the form of its implementational assumptions — is an issue that we consider carefully in the Discussion.

Setting aside this controversy, it is important to note that almost all previous models — from all theoretical frameworks — have investigated relatively straightforward morphological systems (e.g., English past-tense or German/Arabic noun plural marking) in which each input-output mapping reflects only a single distinction (nonfinite stem vs. past tense; singular vs. plural). Indeed, we are aware of only two connectionist studies that have investigated the acquisition of more complex systems. First, the model of Hare and Elman (1995) simulated verb inflection in Early Old English; a system that is much simpler than modern Finnish or Polish, with just three distinct present-tense forms (1sg, 2sg, plural). Because the focus of that study was processes of language change, the model did not attempt to simulate phenomena observed in language acquisition. Second, the model of Mirković et al. (2011) simulated noun marking in Serbian; a highly complex system with three genders (masculine, feminine and neuter), two numbers (singular and plural), and seven cases (nominative, accusative, genitive, dative, instrumental, locative and vocative). Although the model successfully learned the system, and simulated the performance of native-speaking adults tested as part of the study, it was not evaluated on its ability to simulate phenomena observed in child language acquisition.

In the present work, we therefore investigate whether the connectionist approach used in studies of — with the exception of Mirković et al. (2011) — considerably simpler systems can scale up to simulate the acquisition of systems of verb person/number marking in two highly inflected languages: Polish and Finnish. Crucially, the model is evaluated on its ability not only to learn the system and to generalise to unseen items, but also to simulate error phenomena shown by child learners of the relevant language.

Before presenting the model itself (Section 3), we briefly outline the properties of the inflectional systems of present-tense person/number marking in Finnish and Polish that we investigate in the subsequent elicited production studies (Section 2) and the computational modelling work. These two languages were chosen because they are highly inflected languages that have received previous detailed research attention in the domain of inflectional morphology (albeit primarily with regard to nouns rather than verbs in Polish), and are typologically related only very distantly, thus in effect allowing for replication of our elicited production and modelling work across two independent systems. Importantly, the systems also differ with respect to their regularity — Finnish is considerably more regular than Polish — allowing for investigation of whether our findings replicate across two systems that are, on the surface, rather different.

1.3. Properties of the morphological systems under investigation

1.3.1. Finnish

Finnish is a morphologically complex language from the Finno-Ugric group of the Uralic language family. As a highly inflected language, Finnish marks both person and number, resulting in six possible combinations: 1sg, 1pl, 2sg, 2pl, 3sg, and 3pl. Descriptively speaking, each person/number combination is formed by "adding" an affix to the verb stem, and all finite verbs exhibit person/number agreement. Like many other inflected languages, Finnish is a pro-drop language, which means that the pronoun can be omitted in situations in which it is pragmatically inferable. However, 3rd person pronouns cannot be omitted. Four major verb classes can be distinguished in Finnish (see Table 1).

Classes are determined based on alternations in the verb stem — (i) the number of stems a verb has (vowel stem only or both vowel and consonant stem), (ii) which consonant gradation process applies to the verb stem (strengthening or weakening) and (iii) the vowel stem ending (short vowel stems, diphthong/long vowel stems, *-e* stems and 'contraction' verbs). From these major classes, 15 subclasses (Hakulinen et al., 2004) can be distinguished (see Table A.23 in Appendix A) according to particular verb stem endings and other fairly minor stem changes occurring in certain contexts, for example in the past tense.[1]

As can be seen from Table 1, each person/number combination has its own inflectional affix, regardless of the class to which it belongs. The exception is the 3sg context, in which the final vowel in the stem is lengthened unless the stem ends in a diphthong or long vowel. In colloquial Finnish, 3pl forms are mostly produced as 3sg (c.f., Mielikäinen, 1984), and thus, as in Räsänen et al. (2016), the present study did not elicit 3pl forms. Additionally, the 1pl form is often replaced by the passive form in informal contexts, and thus both 1pl and passive responses to 1pl contexts were treated as correct in the experiment.

1.3.2. Polish

Polish is a member of the Slavic group of languages, belonging to the Indo-European family. As in Finnish, finite Polish verbs obligatorily mark person/number with the same six combinations (1sg, 2sg, 3sg, 1pl, 2pl, 3pl), and allow for subject omission. Unlike in Finnish, person/number marking affixes in Polish are different between classes. For present tense imperfective verbs there are four sets of suffixes, yielding four general conjugation classes. The class of a verb is evident from the 1sg and 2sg forms. The other forms of the present tense can be predicted from these two forms. For the respective sets of present tense affixes, see Table 2.

[1] Verbs in the vowel stem-only classes I and II can have both a weak and a strong vowel stem if consonant gradation applies to the consonants in the verb stem. All verbs are also subject to vowel harmony rules, as in the language as a whole. In these classes, the 1st and 2nd person forms take the weak vowel stem, and the 3rd person forms take the strong vowel stem. In the vowel and consonant stem classes III and IV, all person forms take the strong vowel stem.

Table 2
Examples of person/number marking in verbs from each major verb class in Polish, along with its infinitive form and English translation.

Class English	I 'to buy'	II 'to like'	III 'to seek'	IV 'to know how'
inf	kupować	lubić	szukać	umieć
1sg	kupuj-ę	lub-ię	szuk-am	umi-em
2sg	kupuj-esz	lub-isz	szuk-asz	umi-esz
3sg	kupuj-e	lub-i	szuk-a	umi-e
1pl	kupuj-emy	lub-imy	szuk-amy	umi-emy
2pl	kupuj-ecie	lub-icie	szuk-acie	umi-ecie
3pl	kupuj-ą	lubi-ą	szuk-ają	umi-eją

From these major classes, 12 subclasses can be distinguished (Saloni, 1976; Tokarski, 1951). Subclasses in Polish are identified on the basis of (i) stem changes from infinitive to present tense, (ii) stem changes within the present tense paradigm, (iii) stem changes across tenses. These subclasses can in effect be treated as phonological neighbourhoods. The overwhelming majority (75%) of all verbs in Polish belong to class I, group 4 (26%), class II, group 6 (23%) and class III, group 1 (26%). For a more detailed description of the subclasses, see Table A.25 in Appendix A.

2. Elicited production studies

When designing the studies set out below, our goal was to advance the state-of-the-art in elicited-production studies of verb morphology in three ways that elucidate the mechanisms of language acquisition. First, we focussed on highly-inflected languages in which a different form of the verb is used for each present-tense person + number combination. In contrast, most previous studies have focussed on languages such as English, in which a single form is used in a variety of present-tense contexts (e.g., *I/You/We/They play*), and/or on domains in which the verb has (usually) a single correct target form, such as the English past tense (e.g., *played*). The nature of the system has important implications for the type of theoretical account that is viable. For example, theories based on the notion of a morphological default form (e.g., Prasada & Pinker, 1993), clearly apply much more straightforwardly to systems of the English-past-tense type. Second, we devised a new elicitation method which involves using photographs of the participant and experimenter (besides other, unknown, characters), and therefore allows for the elicitation of first, second- and third person (singular and plural) forms. Without this method, it is almost impossible to probe for first person forms without inviting errors caused by the necessity of pronoun reversal; e.g., Experimenter: "In this video, you (2sg)…", Child: "…∗are playing" (2sg), as opposed to the target response "…am playing" (1sg). The method also involves explicitly presenting the child with the pronoun with which to begin her response, leaving no doubt as to the target person + number form required. Third, our goal was to obtain production data that not only constitutes a target for subsequent connectionist modelling but that can, in principle, be used to test directly the predictions of competing theoretical accounts.

In practice, few theoretical accounts are sufficiently well specified to allow for straightforward, uncontroversial derivation of predictions; and even fewer set out their predictions explicitly. However, following Pothos (2005) and Granlund et al. (submitted for publication) we can profitably situate general theoretical approaches along a continuum from those that operate (a) wholly on the basis of formal linguistic rules and categories (e.g., Deen & Hyams, 2006; Hoekstra & Hyams, 1998; Wexler, 1998) versus (b) wholly on the basis of witnessed exemplars (e.g., the exemplar-based and connectionist accounts set out above). In between, lie approaches such as the tolerance principle (e.g., Schuler, Yang, & Newport, 2016), multiple rules (Albright & Hayes, 2003) words-and-rules (e.g., Alegre & Gordon, 1999; Clahsen, Rothweiler, Woest, & Marcus, 1992; Hartshorne & Ullman, 2006; Pinker & Ullman, 2002) and the pre-/protomorphological approach (e.g., Bittner, Dressler, & Kilani-Schoch, 2003; Kenstowicz & Kisseberth, 2014; Stephany & Voeikova, 2009). These approaches posit some sensitivity to frequency and/or learning of phonological regularities, while retaining some role for formal symbolic rules. In general, findings of facilitative effects of token frequency and phonological neighbourhood density are more consistent with approaches towards the exemplar-based end of the continuum; so too are findings that errors are more common for low frequency contexts, and often involve the production of higher-frequency competitors. Findings of excellent performance across the board, irrespective of frequency or phonology, are more consistent with approaches towards the rule-based end of the continuum.

2.1. Experimental method

2.1.1. Participants

Eighty children were tested in Finnish. They were recruited from 9 nurseries located in the Satakunta region in Southwest Finland (with 4 participants tested at home). In Polish, 91 children were tested, and were recruited from 6 nurseries in Warsaw, central Poland and Olsztyn, Northeast Poland (with 10 participants tested at home). All tested children were reported by their parents and teachers to be typically-developing, monolingual speakers of Finnish or Polish. Three Finnish participants and 10 Polish participants were excluded from the final dataset due to failure to provide any scorable responses in the experiment. Thus, the final sample consisted of 77 children in Finnish (46 females; mean age 49.4 months; range 35–63 months) and 81 children in Polish (43 females; mean age 48.7 months; range 35–59 months).

2.1.2. Predictor variables

In Finnish, token frequency counts for each verb form were obtained from the Kirjavainen-Max Planck child-directed speech corpus (Kirjavainen, Kidd, & Lieven, 2017). This corpus consists of 278 files of audio transcriptions of a monolingual Finnish-speaking child interacting with her mother, father and other relatives between the ages of 1;7 and 4;0 (684,000 tokens). Annotations were available for the mother's utterances only. All tokens produced by the mother were included in our analysis. In Polish, token frequency counts were collected from the CDS frequency lists of Haman et al. (2011), which comprise data from seven individual corpora, for a total of 794,000 word tokens in child-directed-speech (directed at children aged between 0;10 and 6;11). From these corpora, all tokens from all adults were included. For the most part, Polish verbs occur in aspect pairs, consisting of 'imperfective' (IPFV) and 'perfective' (PFV) partners, as in *pisać* 'to write' (IPFV) and *napisać* 'to have written' (PFV) (Swan, 2003). In Polish, conjugational endings are the same for imperfective verbs (IPFV) in present tense and perfective verbs (PFV) in future tense; for this reason, in the present study both types of verbs were included in frequency counts.

The previous literature suggests two broad approaches to defining phonological neighborhood density. The simplest and most commonly used approach is the number of "classmates", as defined by a reference grammar (e.g., Dąbrowska, 2008b; Dąbrowska & Szczerbiński, 2006; Kirjavainen et al., 2012; Marchman, 1997; Marchman et al., 1999; Räsänen et al., 2016; Savičiūtė et al., 2018). The precise definition of a class varies from grammar to grammar but, broadly speaking, all members of a class share commonalities in terms of both their base/stem form and their inflected form(s). For example, in the domain of the English past-tense (Marchman, 1997), the *-eep/-ept* class has six members: *weep-wept, keep-kept, sleep-slept, creep-crept, sweep-swept, leap-leapt* (at least in British English). A potential shortcoming of this approach, with regard to the present domain, is that classes are defined by the behaviour of the verb across the person + number marking paradigm as a whole, which may not be relevant in a particular person + number elicitation context. For example, as shown in Table 2, the Polish verbs *kupować*, 'to buy' and *lubić*, 'to like', are classified as belonging to different classes, because — for five out of six person + number contexs – they indeed take different inflectional endings. In 3pl form, however, the target inflection, -ą is the same for both. Thus, when a child participating in the present study is attempting to produce a 3pl form, one can make the argument that it is inappropriate to treat these verbs as members of different classes (depending on the extent to which different person + number forms of a given verb are assumed to be "joined-up" in the learner's system). The second, more complex, approach (e.g., Albright & Hayes, 2003; Granlund et al., submitted for publication) involves calculating a graded similarity metric between the stem and inflected form separately for each target context. For example, *kupować* and *lubić* would be treated as neighbours for 3pl elicitation contexts *kupuj-ą/lubi-ą*, but not 3sg elicitation contexts *kupuj-e/lub-i*. This approach is radically exemplar-based, because it assumes that, when producing a verb form in a particular person + number context (e.g., 3pl), the inflectional behaviour of that verb in all other person + number contexts is irrelevant (and so may be inappropriate, if one rejects a radical exemplar account).

For the present study, we decided to use the simpler class size measure, for three reasons. First, this measure is more objective and transparent, as it involves using "off the shelf" classifications. The continuous measure requires the use of a mathematical model with a number of free parameters, which can be tweaked to reduce or increase any observed effect of PND. Second, the simpler measure allows for more direct comparison with previous studies of Polish and Finnish (Dąbrowska, 2008b; Dąbrowska & Szczerbiński, 2006; Kirjavainen et al., 2012; Räsänen et al., 2016). Third, in a direct comparison of the two approaches, Granlund et al. (submitted for publication) found that the simpler class size measure is more conservative. The use of the more conservative measure is desirable, given that which of the two methods is more appropriate is not merely unclear, but hinges crucially on theoretical perspective. This measure was obtained from reference grammars for Finnish, *Ison Suomen Kieliopin Verkkoversio* (Hakulinen et al., 2004) and Polish, *Słownik Języka Polskiego PWN* (Drabik & Sobol, 2007). In the remainder of this paper, we use the term "class size" to refer to the particular measure used in the present study, and "phonological neighbourhood density" (or "PND") to refer to the concept more generally.

2.1.3. Design

The study used a within-subjects design and an elicited production paradigm. The stimuli consisted of 32 verbs in each language with accompanying videos presented on a laptop computer. All verbs were chosen to be familiar to young children, and suitable for illustration in cartoon animations. For each language, verbs were chosen from 4 major conjugation classes in such a way as to ensure a continuum of low-to-high token frequency within each class. This meant that, for the most part, it was not possible to use translational equivalents across the two languages. As well as the major conjugation classes, we endeavoured to balance verbs across subclasses. For Finnish (see Appendix A Table A.24), we selected eight verbs from each major class, ensuring that this set contained between two and four verbs from each of the 11 subclasses. For Polish (see Appendix A Table A.26), we selected 12 verbs from class I (four each from three subclasses), eight from class II (divided into four per subclass), eight from class III and 4 from class IV (classes III and IV are not divided into subclasses). For each language we selected, as a practice verb in addition to the experimental verbs, a highly frequent verb from the most frequent subclass.

For each language, we created 10 pseudo-randomized lists, each containing 16 of the 32 verbs. Each child completed one list (i.e., half of the total experimental design) in order to minimize fatigue. For each of the 16 verbs assigned to each participant, we attempted to elicit all six present tense forms (1sg, 2sg, 3sg, 1pl, 2pl, 3pl) in Polish, and all but 3pl (which is not commonly used in spoken language) in Finnish, for a total of 96 and 80 trials per participant in Polish and Finnish, respectively.

2.1.4. Procedure

Each child was tested individually in a quiet setting, and completed three sessions: a training/practice session (20–25 min) and two experimental sessions (each 15–25 min), each containing half of the test stimuli for each child (i.e., 48 trials in Polish and 40 in

Finnish). The first experimental session was presented immediately after the training/practice session, with a break of a few minutes. The second experimental session was conducted either after a break of a few hours, or on the next day.

Videos were displayed on a laptop computer (13-in. screen). All animations were presented through Processing 3.0.1, a Java-based program. The experimenter presented each stimulus by pressing a forward button on the keyboard. All actions were continuous or — for inherently telic verbs — performed 1.5 times. This allowed each animation to end on an informative freeze-frame which most clearly depicted the action being performed, and which remained on-screen while the child produced her response. All sessions were audio recorded using Audacity.

The child was seated in front of the laptop computer with a 'talking' toy fox positioned next to the laptop, facing the child and the experimenter. The toy fox's internal speakers were connected to the laptop. The child was told that she would be playing a game with the experimenter in which they would watch and describe some videos of different actions.

In the first part of the training/practice session, the experimenter modelled each of the 16 target verbs assigned to that child — in gerund form in Polish (e.g., *Ten film jest o skakaniu* 'This film is about jumping') and in the 4th infinitive partitive form in Finnish (e.g., *Tässä on hyppimistä* 'This is jumping') — and asked the child to repeat this form after watching the video. In the second part of the training/practice session, the child was again shown each video, and this time asked to label each action without the experimenter's help. This was done to ensure that the child could correctly name all the actions which would be used in the experimental session. All the actions in the training videos were performed by a character unknown to the participant.

All characters depicted in the videos had a pasted-on head (added in real time by the stimulus-presentation software), taken from a photograph of a real person (see Fig. 1). These characters were

- **1sg:** the child (a photograph was taken after the second part of the training/practice session)
- **2sg:** the experimenter
- **3sg:** an adult male, also used in the training/practice sentences with gerund/infinitive forms
- **1pl:** the child and the experimenter
- **2pl:** the experimenter and an unknown 3rd character (the Polish experimenter for the Finnish children, and vice versa)
- **3pl:** (only in Polish) an adult male 2 and an adult female 2

To our knowledge, this method of using photographs of heads that are added in real time is novel and allows researchers, for the first time, to elicit forms such as 1sg and 1pl using video animations. The software code (for Processing), copyright-free video animations and sample head photographs — along with instructions — can be downloaded from https://osf.io/uepz9/.

In the third part of the training/practice session, participants were presented with animations of each of the person/number combinations for the practice verb (in Polish, *kupować*, 'to buy' and in Finnish, *nukkua*, 'to sleep') in randomised order. The child was instructed to listen to "Mr. Fox" as he would say the first word (the pronoun) that they needed to repeat before describing the scene, for example, Fox: *Mää* ('I') – Child: *Mää nukun* ('I sleep-1sg'). The child was instructed to always wait until the fox had produced the first word before producing their response. This part of the procedure (which is also novel compared with previous studies) was designed to ensure that the target person/number context was unambiguous for the child. For the training/practice session only, the experimenter corrected any erroneous responses (or responses lacking the pronoun), and asked the child to repeat the correct target utterance.

The two experimental sessions followed the same procedure as the third part of the training/practice session — with all verb+person/number elicitation contexts presented in random order — except that the experimenter did not provide corrections or other feedback. If the child did not respond on a particular trial, the fox repeated the prompt once. If the child still did not respond,

Fig. 1. A screenshot of an example animation in 3sg context ('to dance').

Table 3
Percent of scorable (correct and incorrect) and unscorable (types of errors) responses produced in the study for each language.

		FI (%)	PL (%)
Scorable	Correct	57.9	71.9
	Incorrect	3.6	3.0
Unscorable	(a) Infinitive or gerund	6.3	1.6
	(b,c) No/unintelligible resp.	6.8	8.7
	(d,e) Only pronoun or stem	3.3	0.07
	(f) Wrong tense	2.6	0.4
	(g) Wrong verb	18.8	12.7
	(h) Wrong pronoun	0.8	1.2

the experimenter moved on to the next trial. Children were rewarded with stickers throughout the experiment, regardless of the responses produced.

2.1.5. Transcription and coding

The experimenter transcribed the participant's responses to each trial online. If a response was unclear, it was marked as such and, following the session, the audio recording was consulted for clarification. In total, there were 7,287 responses in the Polish study, and 5,360 responses in the Finnish study.

All responses were coded as either scorable (1 = correct; 0 = incorrect) or unscorable (NA). The scorable responses were classified as either correct (1; the participant produced the correct person/number marked form of the target verb) or incorrect (0; the participant produced an inappropriate person/number marked form of the target verb). As both Finnish and Polish are pro-drop languages, which allow subject omission, production of the pronoun was not required for a response to be counted as scorable. A response was deemed unscorable if the participant produced (a) an infinitive form, (b) no response, (c) an unintelligible word or affix, (d) a pronoun on its own, (e) a verb stem on its own, (f) a non-present-tense form (e.g., past), (g) a non-target verb, or (h) a non-target pronoun. Table 3 reports the proportions of unscorable responses produced in each language.

Note that, as in Räsänen et al. (2016) and multiple previous studies on verb inflection (e.g., Deen, 2004; Harris & Wexler, 1996), the only responses counted as incorrect were those with erroneous person/number marking on the target verb. This decision was taken for two reasons. First, many generativist-nativst models (e.g., Deen, 2004; Harris & Wexler, 1996; Hoekstra & Hyams, 1998; Legate & Yang, 2007; Smoczyńska, 1985; Wexler, 1998) assume that children pass through a stage in which infinitive forms are licensed by the grammar. Therefore, in order to be conservative with regard to these models, it is necessary to avoid treating such forms as errors. Second, since an important goal of the present study is to directly compare children with a connectionist model, it is necessary to disregard erroneous child forms that are not produced by the model for purely implementational reasons: Because the model is trained to produce person + number marked present-tense forms in response to a stem, it does not produce infinitive, bare-stem or past-tense forms.

As in Räsänen et al. (2016), gradation errors, local dialect forms or consonant misarticulations were not considered errors. However, because stem changes play a large role in determining verb classes, any stem errors, as well as any class errors (responses in which the child produced a person/number ending from a different class, but in the correct person/number context) were counted as scorable errors. To calculate the reliability of the coding of error types, 15% of the responses were transcribed by another native speaker of each language, independently of the original transcriber. Agreements were high (Polish: 98.5%, Finnish: 95.8%).

2.2. Experimental results

2.2.1. Descriptive analysis of errors

Before moving on to the statistical analyses, we first present a descriptive analysis of errors designed to investigate the claims — based on previous findings — that

- Overall error rates are low, but are high for rare person/number contexts (see Section 1.1.1).
- Many errors (usually a clear majority) are of one of three types: (a) frequency-based substitutions; (b) near-miss errors; (c) conjugation-class errors (see Section 1.1.3).

In Finnish, 7.3% of all scorable responses were errors ($N = 240$). Table 4 shows overall error rates in each person/number combination. Most errors were observed in 2pl, which is the rarest context according to CDS input frequencies. Conversely, the most frequent 3sg context showed fewest errors. Over half of all errors involved the replacement of another person/number form with the 3rd singular form (see Table 5), which is — for every verb — the form with the highest frequency in CDS. Similarly, many errors involved the replacement of the 2pl with the passive form, which again is considerably more frequent in CDS. Many of these errors could also be analysed as near miss errors, since they maintain either the person or number of the target (e.g., the production of 3sg forms in 1sg and 2sg target contexts, maintaining singular number). Conjugation class errors, however, are a marginal phenomenon in Finnish, since (as shown in Table 1), only 3sg forms (and passive forms, which were not targets in the experiment) have different

Table 4
Percentage of errors in each person/number combination for Finnish and Polish.

	1sg	2sg	3sg	1pl	2pl	3pl
Finnish (%)	7	7	2	9	11	
Polish (%)	6	10	2	4	17	10

Table 5
The most frequently occurring errors in Finnish, along with the frequency of each error and the proportion of that type of error out of all errors (%). In the 'Example' column, the word in parentheses represents the correct form. The final column shows whether the error can be attributed to the use of a more frequent person/number combination (F). Only errors in which $n > 5\%$ were included.

Error	Type	Count	%	Example	Sub.
3sg instead of 2pl	P/N	40	16.7	kävelee (kävelette)	F
3sg instead of 1sg	P/N	39	16.3	ajattelee (ajattelen)	F
3sg instead of 2sg	P/N	33	13.8	lukee (luet)	F
3sg instead of 1pl/PASS	P/N	18	7.5	piirtää (piirretään)	F
PASS instead of 2pl	P/N	18	7.5	pudotaan (putoatte)	F
Adding + tse in PASS in IVd	Stem	14	5.8	valitsetaan (valitaan)	
PASS: strong, not weak, stem	Stem	13	5.4	nousetaan (noustaan)	
Total		193	73		

inflectional endings (e.g., *nukku*-**u**, *saa*-**0**, *tule*-**e**, *hyppää*-**0**) for each conjugation class (c.f., 1sg forms *nuku*-**n**, *saa*-**n**, *tule*-**n**, *hyppää*-**n**).

In Polish, 7.8% of all scorable responses contained errors ($N = 432$). As Table 4 shows, error rates in Polish also patterned according to the CDS frequency of person/number combinations, with — again — the highest and lowest error rates observed in 2pl and 3sg contexts, respectively. Again, a large proportion of errors reflected the replacement of lower frequency target forms (e.g., 2pl, 2sg) with a higher frequency form of the target verb (1pl, 3pl, 3sg); many of which are also analysable as near-miss errors. Unlike for Finnish, the most frequently occurring error was a conjugation-class error (i.e., overgeneralisation); specifically, the use of class III instead of class I-9 (see Table 6). This phenomenon appears to reflect phonological neighbourhood density/type frequency: Since class III is much larger than subclass I-9, the former contains many stems that are phonologically similar to a given stem in class I-9, but actually take a different pattern of inflections (sometimes known as "enemies", e.g., Marchman, 1997; Marchman et al., 1999, as opposed to "neighbours" or "friends"). In sum, only one of the most frequently occurring error types (the use of 1pl in 2sg contexts) does not constitute a frequency-based substitution, near-miss error or conjugation-class error.

In summary, for both Finnish and Polish, this error analysis is highly consistent with the findings of previous studies that overall error rates are low, but are high for rare person/number contexts (see Section 1.1.1), and that most errors constitute (a) frequency-based substitutions, (b) near-miss errors, or (c) conjugation-class errors (see Section 1.1.3). Both of these findings are therefore clear targets for simulation in the subsequent modelling work.

2.2.2. Statistical analysis of input-based predictors

These analyses were designed to investigate the claims – based on previous research – that

- Errors are more common for individual inflected forms with low token frequency, which has considerable support from previous studies (albeit mostly outside the domain of person/number marking) (see Section 1.1.2).
- Errors are less common for verbs that score high on phonological neighbourhood density (PND)/type frequency/islands of reliability/class size, which also has considerable support from previous studies (albeit again from other domains) (see Section 1.1.4).
- Phonological neighbourhood density (PND) effects may be smaller for forms with higher token frequency, which has weak support

Table 6
The most frequently occurring errors in Polish, along with the frequency of each error and the proportion of that type of error out of all errors (%). In the 'Example' column, the word in parentheses represents the correct form. The final column shows whether the error can be attributed to the use of a more frequent person/number combination (F) or to the use of overgeneralisation (O). Only errors in which $n > 5\%$ were included.

Error	Type	Count	%	Example	Sub.
Class III instead of class I	Class	77	18	mazasz (mażesz)	O
1pl instead of 2pl	P/N	59	13	wieziemy (wieziecie)	F
3pl instead of 2pl	P/N	43	10	tańczą (tańczycie)	F
1pl instead of 2sg	P/N	27	6	gnieciemy (gnieciesz)	
3sg instead of 2sg	P/N	26	6	poluje (polujesz)	F
Total		232	53		

from a single study (Räsänen et al., 2016) (see Section 1.1.5).
- Effects of phonological neighbourhood density and token frequency may decrease with age. The former has weak support from Räsänen et al. (2016); the latter has no support from previous studies of which we are aware (see Section 1.1.6).

We tested for these effects using linear regression, and additionally — mainly due to convergence problems with these frequentist models — computed Bayesian posterior distributions. We explain the methods of both approaches and the differences between them below. In order to account for both item and participant variance, the frequentist analysis was conducted using logistic linear mixed-effects regression models (Baayen, Davidson, & Bates, 2008) with the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2016). We fitted a generalised mixed-effects model with a binomial link function, with response accuracy (incorrect, correct) for scorable trials only as the dependent variable (0, 1) and age (in months), target token frequency and class size as fixed effects. For each language separately, the predictors of token frequency and class size were log-transformed, centred and standardised (divided by their standard deviations), while age was centred to 48 months for both languages and then standardised. All two-way interactions were included in the model (see Sections 1.1.4–1.1.6). Three-way interactions were not included because no predictions were made regarding three-way interactions, and our design is underpowered for their detection. Random intercepts were specified by participant and verb, and random slopes for all fixed effects (except age) by participant. Maximal models were run initially and, in cases of non-convergence, the random structure was simplified using the procedure outlined in Barr, Levy, Scheepers, and Tily (2013). P-values were obtained using ANOVA model comparison (likelihood ratio test).

In addition to the frequentist analysis, we fitted Bayesian generalised linear mixed models with a binomial link function using the `rstan` package (Stan Development Team, 2015) with the `rstanarm` extension (Gabry & Goodrich, 2016). The use of Bayesian models has several advantages (see Nicenboim & Vasishth, 2016 for an overview). One major advantage is that a maximal random effects structure (Barr et al., 2013) can be fitted without convergence problems. Another advantage of Bayesian inference is that it allows the computation of *credible intervals*, which provide the range within which the true effect lies with a certain probability given the data. Thus, credible intervals have a more straightforward interpretation than the often-misunderstood confidence intervals (see Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). As explained in more detail below, the same is true for so-called "Bayesian P-values", which — unlike frequentist p-values — allow direct probabilistic statements about the effect in question, given the data, without reference to a null hypothesis. Despite their advantages, there is still hesitation in the field about fully adopting Bayesian methods. We therefore report both the frequentist and Bayesian results, and base our conclusions on both of them.

In the Bayesian analyses, we always fitted maximal models as justified by the design (random intercepts for participant and verb, and by-participant random slopes). We used weakly informative priors for fixed and random effects (i.e., we did not impose any prior information on the estimates).[2] We report the mean estimate β, the lower and upper limits of the 95% credible interval, and the probability of the effect being smaller than (for negative estimates) or greater than (for positive estimates) zero — in both cases abbreviated as P_β. Note that the probability P_β in the Bayesian sense can be interpreted literally as the probability of the true effect being smaller/greater than zero, given the data. Thus, P_β here is fundamentally different from the *p-value* in the sense of null hypothesis significance testing (in NHST, the p-value is the probability of an effect of at least the observed magnitude, given that the null hypothesis is true).

As there is no binary decision threshold for significance in the Bayesian approach, we interpret the Bayesian results as follows:

- If $P_\beta \geqslant 0.95$ or the 95% credible interval does not span zero, we interpret this as strong evidence for an effect given the data.
- If the credible interval contains zero but the probability P_β is relatively high (roughly around 0.85 or higher), we say that there is weak evidence.
- If P_β is close to 0.5, we conclude there is no evidence for an effect.

It is important to note that using both frequentist and Bayesian approaches does not amount to so-called *p-hacking*. This approach can be considered more conservative as we use two different computational methods for fitting the same linear models on the same data and base our conclusions on the results of both methods in combination.

In the frequentist models for both languages, even the simplest random structure (by-participant and by-verb random intercepts) failed to converge. Therefore, we removed the age × token frequency and age × class size interactions from the model, resulting in a converging model.

In Finnish, the frequentist analysis showed significant main effects of class size, token frequency, and age (see Table 7) — the positive beta values indicate that accuracy improved with larger class sizes, with higher token frequency and with age. The interaction between token frequency and class size was not significant and was not included in the final model (the effect is still shown in the table for completeness). The Bayesian analysis supports all the main effects, showing strong evidence for all three. In addition, the Bayesian analysis showed weak evidence for a token frequency and age interaction, suggesting that the influence of token frequency on accuracy decreases with age.

In Polish, the frequentist analysis showed significant main effects of both class size and token frequency; again, children's accuracy increased with greater phonological class size and with higher token frequency. Neither age nor either of its interactions reached significance. They were therefore not included in the final model. As for Finnish, the interaction between token frequency

[2] The priors for the intercept and slope were a Student t-distribution with 2 degrees of freedom and mean 0. For the random effects correlation matrix, a so-called LKJ prior was used (see Sorensen, Hohenstein, & Vasishth, 2016, for a tutorial).

Table 7
Frequentist and Bayesian analysis results for Finnish. For the Bayesian analysis, the estimated mean is reported as well as the lower and upper limits of the 95% credible interval and the probability of the effect being <0 (for negative coefficients) or >0 (for positive coefficients) given the data. Significant effects (frequentist) or strong evidence (Bayesian) are marked in bold.

	Frequentist analysis				Bayesian analysis			
	Est.	SE	$\chi^2(1)$	p	Mean	Lower	Upper	P_β
Intercept	3.09	0.24	–	–	3.40	2.98	3.84	1
ClassSize	**0.17**	**0.065**	**5.78**	**0.016**	**0.15**	**−0.01**	**0.30**	**0.97**
TokenFreq	**0.38**	**0.064**	**39.14**	**< 0.001**	**0.44**	**0.28**	**0.62**	**1**
Age	**0.082**	**0.028**	**8.71**	**0.0032**	**0.07**	**0.0009**	**0.14**	**0.98**
ClassSize:TokenFreq	0.021	0.031	0.45	0.50	0.01	−0.06	0.08	0.63
ClassSize:Age	–	–	–	–	0.02	−0.01	0.02	0.59
TokenFreq:Age	–	–	–	–	−0.01	−0.03	0.01	0.86

and class size was not significant. The Bayesian model showed strong support for the class size and token frequency main effects, and also indicated some evidence for a positive effect of age. As for Finnish, the Bayesian analysis in Polish also showed weak evidence for a negative interaction of token frequency with age (see Table 8).

In both languages, the Bayesian analysis showed weak evidence for a negative interaction of token frequency with age with a $P(\beta < 0)$ of 0.86 and 0.88. This might suggest that a study with more data would show strong evidence that the effect of token frequency decreases with age. We, therefore, pooled the data from both languages for an exploratory analysis to see whether there would be stronger evidence for the token frequency × age interaction or even the token frequency × class size interaction. The pooled exploratory analysis used the Bayesian method only. We used the same model structure as in the confirmatory analyses above, with the addition of a random intercept for language.

The results are shown in Table 9. This model showed slightly stronger evidence for an interaction of token frequency and age, the probability of the effect now being 0.91. The evidence for a class size × token frequency interaction, however, remains very weak, and the estimate is greater than zero, which is the opposite of the predicted direction. All other effects remained the same as in the previous analyses or became stronger.

2.3. Discussion of the experimental results

For both languages individually, and for the pooled analysis, the statistical analysis strongly supported the claims made on the basis of previous studies that errors are more common for individual inflected forms with low token frequency (Section 1.1.2) and less common for verbs with high phonological neighbourhood density (PND, Section 1.1.4). However, contra Räsänen et al. (2016), the present study yielded no evidence for the possibility that phonological neighbourhood density (PND) effects may be smaller for forms with higher token frequency (Section 1.1.5; indeed, the estimate for the interaction was never in the predicted direction). However, in line with Räsänen et al. (2016), the present study yielded some weak evidence (which was slightly stronger in the pooled analysis) for the possibility that the effect of phonological neighbourhood density may decrease with age (Section 1.1.6).

When these results are combined with those of the descriptive error analysis (Section 2.2.1) we are left with a very clear picture of the true pattern of correct usage and error that any account of the acquisition of present tense person/number marking morphology must be able to explain:

(a) Overall error rates are low, but are high for rare person/number contexts.
(b) Most errors constitute (a) frequency-based substitutions, (b) near-miss errors, or (c) conjugation-class errors.
(c) Effects of both token frequency and phonological neighbourhood density are robust, but there is no evidence for an interaction

Table 8
Frequentist and Bayesian analysis results for Polish. For the Bayesian analysis, the estimated mean is reported as well as the lower and upper limits of the 95% credible interval and the probability of the effect being <0 (for negative coefficients) or >0 (for positive coefficients) given the data. Significant effects (frequentist) or strong evidence (Bayesian) are marked in bold.

	Frequentist analysis				Bayesian analysis			
	Est.	SE	$\chi^2(1)$	p	Mean	Lower	Upper	P_β
Intercept	1.16	1.21	–	–	1.48	−0.86	3.84	0.89
ClassSize	**0.23**	**0.07**	**9.46**	**0.002**	**0.22**	**0.06**	**0.38**	**1**
TokenFreq	**0.27**	**0.05**	**30.31**	**< 0.001**	**0.26**	**0.16**	**0.36**	**1**
Age	0.04	0.02	2.41	0.12	0.03	−0.02	0.08	0.91
ClassSize:TokenFreq	0.01	0.02	0.49	0.48	0.01	−0.03	0.05	0.77
ClassSize:Age	–	–	–	–	−0.0003	−0.01	0.01	0.5
TokenFreq:Age	–	–	–	–	−0.008	−0.02	0.006	0.88

Table 9
Bayesian analysis results for pooled model including both Finnish and Polish. The table reports the estimated mean as well as the lower and upper limits of the 95% credible interval and the probability of the effect being <0 (for negative coefficients) or >0 (for positive coefficients) given the data.

	Mean	Lower	Upper	P_β
Intercept	3.14	2.18	4.11	0.997
ClassSize	**0.20**	**0.08**	**0.32**	**1**
TokenFreq	**0.34**	**0.24**	**0.44**	**1**
Age	**0.07**	**0.03**	**0.11**	**1**
ClassSize:TokenFreq	0.02	−0.02	0.06	0.83
ClassSize:Age	0.002	−0.01	0.01	0.65
TokenFreq:Age	−0.01	−0.02	0.004	0.91

between these factors.

(d) Although children's performance increases with age (as would be predicted under any account), there is only weak and no evidence respectively that effects of token frequency and phonological neighbourhood density decrease with age. That is, we currently have little evidence for the proposition that children's learning changes qualitatively with age.

In terms of the theoretical continuum outlined above (Section 2), these findings would seem in general to be more consistent with theories towards the exemplar- than rule-based end of the continuum. Although the relatively low overall error rates observed are consistent with purely rule-based accounts, the fact that such errors nevertheless pattern according to frequency and phonological neighbourhood density is difficult to explain under accounts that posit only formal rules and categories. With the set of phenomena that characterize children's learning of present-tense person/number marking morphology clear, we now turn to our attempts to build a computational model that simulates these phenomena.

3. Computational model

3.1. General architecture

The goal of the simulations was to test whether a connectionist model of the kind used for learning the English past tense previously can scale up to the complex inflection paradigms of Finnish and Polish, and whether it exhibits similar error patterns to the children's when trained on a similar task — i.e., the production of inflected verb forms on the basis of child-directed speech input. We therefore trained artificial neural network models on a mapping from verb stems to inflected forms using a standard three-layer architecture with 200 hidden units and phonological form representations on the input and output layers. Networks of this kind have been used for modelling the English past tense in MacWhinney and Leinbach (1991), Plunkett and Marchman (1991, 1993), Daugherty and Seidenberg (1992) and Plunkett and Juola (1999), and also for modelling the Arabic (Plunkett & Nakisa, 1997) and German (Hahn & Nakisa, 2000) noun plural system. A schema of the model architecture used here is shown in Fig. 2.

The *hidden* units in a multi-layer network — so called because they do not directly receive input or produce output — introduce a non-linear component into the mapping process by allowing for arbitrary internal representations of the input. Generally, a large number of hidden units provides the network with more memory to reproduce trained examples, while a lower number of hidden units forces the network to form more abstract representations, which increases its ability to generalise to untrained examples. The majority of previous modelling literature concerned with learning inflection systems found 200 hidden units to be a good compromise between memory and generalisation (Daugherty & Seidenberg, 1992; MacWhinney & Leinbach, 1991; Plunkett & Juola, 1999; Plunkett & Nakisa, 1997). We found this number to work well for our model, too (simulations with higher and lower numbers were carried out, but no major differences were observed apart from differences in speed of learning).

There are, however, differences between previous models and our simulations. For example, in order to increase performance, MacWhinney and Leinbach (1991) used a more complex architecture with two hidden layers of 200 units each and additional connections that directly copied the input stem to the output layer. Plunkett and Marchman (1991, 1993) used artificially created words to train the model, and Plunkett and Marchman (1993) used a training regime that incrementally increased the corpus size

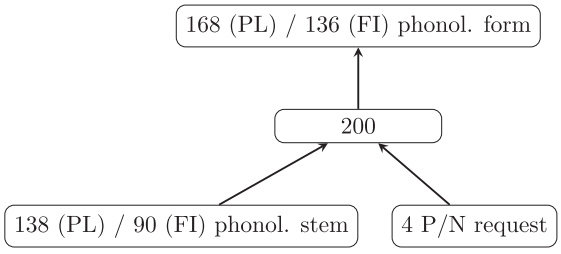

Fig. 2. Model architecture.

during training. Most importantly, none of these models were trained on child-directed speech corpora and directly compared to children's elicited production performance. Furthermore, as already mentioned, we only know of one model of a similar kind that learned an inflection system of similar complexity to Finnish or Polish verb marking — namely Mirković et al. (2011) on Serbian noun case and gender marking. However, like the other models, this model was not analysed with respect to developmental observations. In addition Mirković et al. (2011) used a more complex architecture with two hidden layers and recurrent connections to the output layer in order to produce multi-syllabic words in a mono-syllabic output frame. While multiple hidden layers or an incremental training regime may increase performance, the aim of our simulations was to assess a minimal model on its ability to learn complex inflection.

As input, the model received the stem of the 4th infinitive partitive form in Finnish (which was also used in the training sessions of the experiment) and the stem of the 3sg form in Polish. The decision to map from stem to inflected form is largely one of implementational convenience; we do not assume that children start out with a bare stem, to which they subsequently apply an inflection. Indeed, this seems extremely unlikely given that, for both languages (unlike, for example English), bare stem forms are never encountered in the input. Rather, we assume that children are sensitive to the association between the lexical semantics of a verb (e.g., *sleeping*) and the various phonological forms they hear in the input (e.g., Finnish *nukun*, *nukut*, *nuku* etc.), such that the lexical semantics (e.g., *sleeping*) activates all of these forms. Under this assumption, presenting the "stem" means nothing more than presenting the common phonological material that is shared by all forms cued by the relevant lexical semantics. An alternative would be a pseudo-semantic, *localist* representation where each verb lemma is encoded by an input node (Hare & Elman, 1995; Joanisse & Seidenberg, 1999; Mirković et al., 2011). In order to check that our findings do not hinge crucially on our implementational decision, we also run supplementary models that use a localist representation, and do not receive stems as input.

3.2. Input and output representation

The input layer comprised 94 binary units in Finnish and 142 in Polish, four of which were used to encode person and number of the requested inflection (also see Fig. 2). Three of these units were used for person (first/second/third, each represented by a single unit that could be on or off), and a single unit for number (singular/plural). For example, the second person plural would be encoded as `0101`. Discretely representing person and number in the input is standard in all previous models of the kind and does not necessarily entail the idea of person and number categories being innate in a linguistic sense. Rather, it need entail — minimally — only the assumption that children are sensitive to the semantics of person and number; cues that are clearly visible in the videos seen by the children in the present study. We come back to this point in the General Discussion.

The remaining input units represented the phonemes of the stem, using right-justified syllable templates of `CCVVCCVVCCVV` for Finnish and `CCCVVCCCVVCCCVVCCC` for Polish, where `C` and `V` stand for consonant and vowel slots, respectively. The template structure was chosen for each language such that the majority of verb stems from the CDS corpora could be fitted. Syllable templates have been widely used (Daugherty & Seidenberg, 1992; Hahn & Nakisa, 2000; Joanisse & Seidenberg, 1999; MacWhinney & Leinbach, 1991; Mirković et al., 2011; Plunkett & Juola, 1999; Plunkett & Nakisa, 1997) and are a method of aligning words of different structure and length in order to make their (dis-) similarities available to the network. The alignments of five example stems in Finnish are shown in Table 10. For empty slots, all their units' activation was set to zero.

It is important to acknowledge that, by right-justifying verb representations, we are in effect assuming that children know — or have already learned — that the ending of the stem or the form is privileged over the beginning in terms of determining verbs' inflectional properties (e.g., that children follow something like Slobin's (1973) operating principle "pay attention to the ends of words"). This is psychologically plausible, given that a recency effect (i.e., a recall advantage for more recently presented material) is well established as a robust psychological phenomenon (e.g., Murdock, 1962). Right-justified template representations were also used in MacWhinney and Leinbach (1991), Plunkett and Juola (1999) and Hahn and Nakisa (2000).

Each of the `C` and `V` slots represented one phoneme using eight (for consonants) or seven (for vowels) binary units to encode manner, place, voicing and length of articulation. Tables 11 and 12 show the encoding schemes for consonants and vowels, respectively — an extension of the "Pat-Pho" scheme by Li and MacWhinney (2002), which represents phonemes in a compact distributed way that preserves feature similarity in the representation. The resulting binary codes for all phonemes used in the modelling are provided in Tables B.27 and B.28 in Appendix B. On the output layer, the stem templates were extended to accommodate the endings. For Finnish, the extension consisted of a `CCVVCC` pattern, and for Polish of `VCVC`. As a result, the output layer consisted of 136 units for Finnish and 168 for Polish.

Table 10
Example of template alignments for five Finnish input stems.

C	C	V	V	C	C	V	V	C	C	V	V
–	–	–	–	–	–	–	a	–	s	–	u
–	–	–	–	–	–	–	a	s	t	–	u
–	–	–	–	–	h	a	i	h	t	–	u
–	–	–	–	–	h	e	i	–	l	–	u
–	h	–	e	r	m	–	o	s	t	–	u

Table 11
Consonant scheme used for encoding phonetic features in Polish and Finnish models in eight units (based on "Pat-Pho", Li and MacWhinney, 2002).

	1		2	3	4		5	6	7		8
Voiced	0	Bilabial	0	0	0	Nasal	0	0	1	Short	0
Voiceless	1	Labio-dental	0	0	1	Stop	0	1	0	Long	1
		Dental	0	1	0	Fricative	1	0	0		
		Alveolar	0	1	1	Approximant	0	1	1		
		Palato-alveolar	1	0	0	lateral	1	1	0		
		Palatal	1	0	0	Trill	1	0	1		
		Velar	1	1	0	Affricate	1	1	1		
		Glottal	1	1	1						

Table 12
Vowel scheme used for encoding phonetic features in Finnish and Polish models in seven units (based on "Pat-Pho", Li and MacWhinney, 2002).

	1	2	3		4	5	6		7
Front	0	1	0	High	0	1	0	Short	0
Central	1	1	0	High-nasal	0	1	1	Long	1
Back	1	0	0	Mid	1	1	0		
Front-rounded	0	1	1	Mid-nasal	1	1	1		
Central-rounded	1	1	1	Low	1	0	0		
Back-rounded	1	0	1	Low-nasal	1	0	1		

3.3. Training corpus and predictor variables

Finnish. From the Finnish child-directed speech corpus (Kirjavainen et al., 2017), we extracted all present tense verb forms in the five person/number contexts 1sg, 2sg, 3sg, passive and 2pl. In Finnish, the passive form is used instead of 1pl colloquially more frequently than the actual 1pl form itself (which is why passive was accepted as correct in the 1pl context in the experiments above). We therefore used just passive forms for the modelling and excluded all 1pl forms (which were extremely rare in the corpus). Conjugation class information was added to the dataset (for evaluation purposes only; this information was not available to the model) according to Hakulinen et al. (2004). The resulting word list contained 2307 unique verb forms. Of these unique forms, 806 were ambiguous between being a present tense verb form and other parts of speech. In order to maximize corpus size, we used ambiguous forms in their present tense verb interpretation and interpolated their token frequency using known frequencies of the respective lemma and person/number contexts. The final Finnish word list contained 1,947 unique forms from 907 verbs with CDS token frequency information. Of these, only 66 verbs were represented in all five person/number contexts that were tested in the experiment (1st, 2nd, and 3rd singular, passive, and 2nd plural). All forms were phonologically transcribed by automatically applying context-sensitive grapheme-to-phoneme rules that we mainly based on the description of Finnish phonology in Suomi, Toivanen, and Ylitalo (2008). Finally, all verb forms in the corpus were segmented into stems and suffixes. As the default input stem for the model, we used the 4th infinitive partitive (strong vowel stem) without the ending *mista*.

Vowel harmony between stems and endings was neutralised in the model input where possible. The Finnish system of vowel harmony applies not just to verb inflection but to the whole language. The acquisition of this system therefore happens very early and children do not show any difficulty with it. The neural network models, on the other hand, would need to learn this fundamental property of the Finnish language at the same time as acquiring verb morphology. As we were interested only in the latter, we removed suffix alternations due to vowel harmony in the training corpus (e.g., all passive suffixes ending in /æːn/ or /Aːn/ were represented as /ɑː n/).

Polish. For the Polish model, we extracted 2880 present tense verb forms from the Haman et al. (2011) Polish child-directed speech frequency list, annotated with person, number and (for model evaluation purposes only) conjugation class. All forms were segmented and the 3sg stem was used as input to the model. For this, 311 3sg forms were added manually. The final corpus contained 3227 unique forms from 1328 verbs in four conjugation classes and up to six person/number contexts. Only 83 verbs were represented in all six person/number combinations. All forms were unambiguous. As for Finnish, the Polish corpus was phonologically transcribed.

As in the analysis of the elicited-production experiments, phonological neighbourhood density was measured as class size, i.e., as the number of verbs (types) in each subclass according to the reference grammar.

3.4. Training and testing procedure

Simulations were carried out with `lens` ("light, efficient network simulator"), created by Douglas Rohde (version 2.64 available from http://web.stanford.edu/group/mbc/LENSManual), running on a Linux machine. At each training trial, one of the verb forms was selected probabilistically according to its token frequency in the CDS corpus. The verb stem (e.g., FI: /roikːu/; PL: /rʃsuj/) was presented on the input layer together with the code for the target person/number context. The input activation was spread through

full connectivity to the units of the hidden layer and from there to the output layer. The output unit activations were then compared to the target activations with respect to the correct form (e.g., FI: /roikut/; PL: /rɨsujɛʃ/ for 2sg). Based on the error, the connection weights between the layers were then adjusted using back-propagation (Rumelhart, Hinton, & Williams, 1988).

Normal training and testing. In each language, the training corpus consisted of 800 randomly sampled verbs, resulting in 1784 forms in Finnish and 2431 in Polish. In order to ensure the generalisability of the results, we trained ten networks for each language, which differed in their initial connection weights (set randomly between −0.1 and +0.1). The learning rate (which scales the weight updates during back-propagation) was set to 0.1, and the momentum (the contribution of the previous weight change to the current weight change) to 0.9. We tested the networks on both the full training corpus and a subset consisting of just the 32 verbs that the children were tested on in the elicited-production experiments (160 forms in Finnish and 192 in Polish). Below, we will refer to these two sets of items as the *training set* and the *test set*, respectively. For the error analysis, the model output was mapped onto the closest matching phonemes using Euclidean distance between the output activations in each phoneme group and all possible target phonemes with respect to the language. Every model output thus consisted of existing phonemes of the target language. An output form was scored as correct when all output phonemes were correct.

Assessing generalisation. The ability to generalise to new items reflects the abstract representations that the model builds from regularities it picks up in the input. As a systematic test of the model's ability to generalise, we trained ten additional networks while withholding one form of each of the 32 test verbs from the training corpus. The form that was withheld for each verb was a different one for each network. For example, network 1 saw all forms of the verb *to jump*, except the 1sg form *hyppään*, while network 2 saw all forms of *to jump*, except 2sg, *hyppäät*, and so on. As a result, each network had a different set of 32 forms missing from the training set. That way, we could assess the generalisation ability for each form of the 32 test verbs.

3.5. Simulation results

3.5.1. Finnish

General performance. All trained forms could be produced correctly after about three million training trials, thus demonstrating that a model without formal symbolic rules is indeed capable of learning a highly complex system of person/number marking. Fig. 3 shows that the model achieved high levels of accuracy more quickly for suffixes than stems. This reflects the fact that Finnish inflectional suffixes are very regular, while there are many difficult stem alternations, often due to Finnish consonant gradation. In fact, when we looked only at the 160 test items used in the experiment above, 100% of suffixes were produced correctly after a total number of only 250,000 training trials.

Generalisation. The Finnish model could correctly generalise 85% of the test forms on average at the end of training. This shows that the model actually generalised across examples instead of rote-learning each form. Most generalisation errors occurred in the passive and involved either not changing the stem correctly (e.g., from /putoɑ/ to /pudotɑːn/) or using a suffix from a different inflection class (e.g., /sɑːtɑːn/ instead of /sɑːdɑːn/).

Fig. 4 shows the Finnish model's performance on the 160 test items when they were either trained or new. Stems and suffixes are plotted separately for 11 inflection classes, arranged by their size (number of verbs per class). The figure shows three main findings. First, the model's production of the correct suffix was almost unaffected by whether or not it had seen that particular item before, indicating extremely successful generalisation. This is due to the fact that Finnish present tense verb inflection is very regular and almost identical across classes. Only in class II were some endings generalised incorrectly. In class II, all verbs build their 3sg form without an overt inflection, while in most other classes the final stem vowel becomes lengthened in 3sg whenever it is not a long vowel already. The difficulty here is that many verb stems in class II end in diphthongs rather than long vowels. It is much easier for the network to recognise a long final stem vowel in the input (lengthening status is encoded in one specific bit for every vowel) and hence determine that no extra suffix is needed. In order to recognise a diphthong, however, a sequence of two phonemes has to be taken into account. Second, stems were particularly hard to generalise in class IV, which is relatively small but has a number of

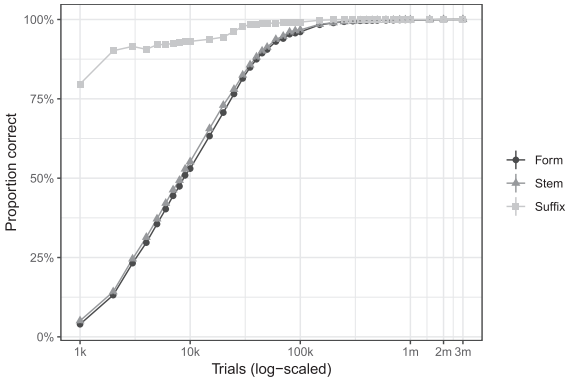

Fig. 3. Finnish model: Proportion of correctly produced stems, suffixes and full forms over training (means and standard errors, averaged across ten runs).

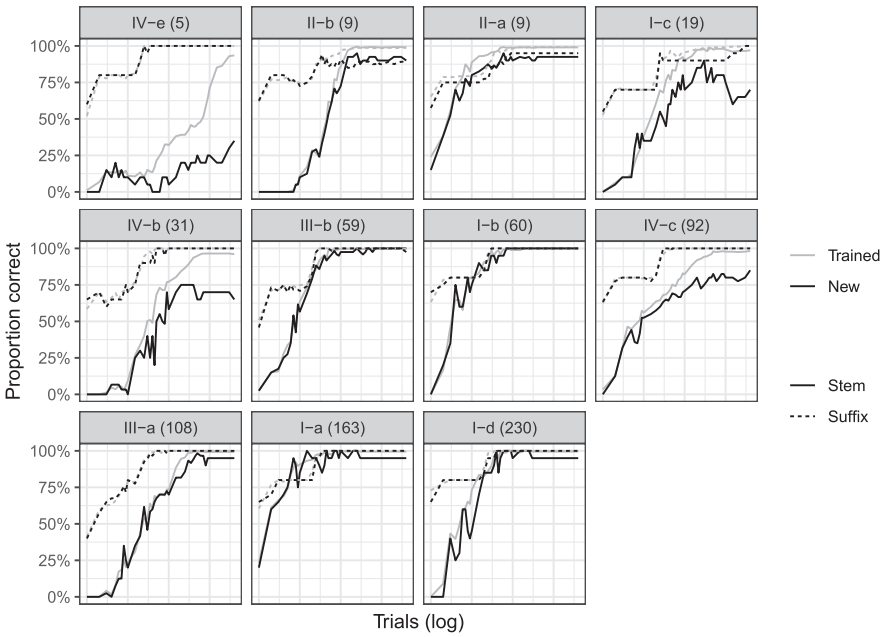

Fig. 4. Finnish model: Stem and suffix accuracy on 160 test items when they were part of the corpus (trained) vs. when they were withheld from the corpus (new) items (averaged across ten runs; numbers in parentheses are inflectional neighbourhood sizes with respect to the training corpus).

different stem endings. Third, Fig. 4 seems to suggest a facilitative effect of phonological neighbourhood density (class size) for both trained and new verbs; a possibility that we investigate statistically in Section 3.5.1.

Descriptive analysis of errors. As for the experimental results, we present a descriptive analysis of errors designed to investigate whether two finding of the elicited production studies are simulated by the model:

- Overall error rates are low, but are high for rare person/number contexts (see Section 1.1.1).
- Most errors are of one of three types: (a) frequency-based substitutions; (b) near-miss errors; (c) conjugation-class errors (see Section 1.1.3).

First, it is clear that, as in the experimental studies, error rates generally pattern by frequency of the relevant person/number context. Like children, the Finnish model showed low error rates in high-frequency contexts, such as 3sg, and higher error rates in low-frequency contexts, such as 2pl. Fig. 5 plots the model's suffix accuracy over training for each person/number context.

The highly frequent 2sg and 3sg forms were acquired fastest. In fact, both displayed over 90% accuracy after only 1000 trials, and 2sg reached 100% accuracy after just 3000–4000 trials. 1sg — another highly frequent context — reached over 90% accuracy after 3000 trials. In contrast, due to the low frequency of the relevant forms, 2pl was not acquired at all until around 15,000 trials, but was then learned very quickly. Although passive forms occur more frequently in the input than 2sg and 1sg forms, they took longer to learn. The reason is that the passive inflection is less regular than the other inflections; a phenomenon that we investigate in more

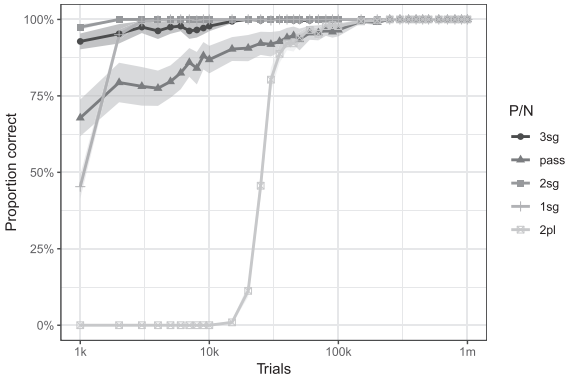

Fig. 5. Finnish model: Development of suffix accuracy in test items by person/number target context over training (means and standard errors, averaged across ten networks). Legend ordering (top to bottom) and line shading (dark to light) represent P/N corpus frequency from high to low.

detail below. Fig. 5 nicely shows that the passive is the only person/number context with a slowly rising learning curve, while all others are very steep.

Although the model generally simulated children's pattern of error rates across different person/number contexts, there is one important exception: Despite its low frequency in the input, 1sg is produced very accurately by children; a finding that the model does not simulate. One possibility is that this phenomenon has a pragmatic explanation: Children prefer to talk about themselves rather than their conversational partners or third persons, and are therefore highly practised in producing 1sg forms, and perhaps also more sensitive to them in the input. The same pattern was found in the naturalistic study of Aguado-Orea and Pine (2015) of verb marking acquisition in Spanish (see also Theakston, Lieven, Pine, & Rowland, 2005, for a corpus study showing a disproportionately late emergence of 2sg compared to 1sg pronoun + auxiliary combinations in English). However, it is important to acknowledge that this pattern was not specifically predicted in advance; this is simply post hoc speculation. Whether or not this explanation is correct, the model cannot, of course, simulate this phenomenon, as its input distribution is based solely on frequencies in child-directed (but not child's own) speech and it is equally sensitive to all of the forms in the paradigm.

Next, we investigated whether — as in the elicited production experiments — most of the model's errors constitute (a) frequency-based substitutions, (b) near-miss errors or (c) conjugation-class errors. However, since a large proportion of near-miss errors cannot be distinguished from frequency-based substitutions, we classified errors only as frequency-based substitutions, conjugation-class errors, or other (random).

Both of the first two error types were indeed common. Frequency-based substitutions appeared mostly early in training (before 10,000 trials) while the number of class errors peaked later, at around 100,000 trials. In order to compare the model to children with regard to the relative frequency of these different error types, it is necessary to examine the model at an intermediate training state (because it eventually reaches 100% accuracy across the board). In order to ensure sufficient variation across items, we chose a point with a relatively high error rate, i.e., at 1000 trials when the mean accuracy for suffixes in 1sg, 2sg, 3sg and passive was at 75% (range: 51–97%) and for suffixes in 2pl still at 0% (the acquisition of 2pl is so far delayed that 1sg, 2sg and 3sg were already at 100% at the time when the model started picking up 2pl). We chose not to test the model at an overall target form accuracy of 75% (60,000 trials) as, at this point, suffixes already had an accuracy between 98 and 100% while the model was still learning the stem alternations.

Fig. 6 shows the distribution of error types after 1000 trials across person/number targets for the 160 test items (averaging across all ten networks). As for children, frequency-based substitutions were common for 1sg and 2pl targets (see Table 13), which were generally replaced by 3sg and passive forms, respectively. Like children, the model showed conjugation class errors (see Table 14) for 3sg forms, and also for passives, which were not targets in the elicited production study (recall that conjugation-class errors are not possible for other person/number forms, which take the same ending across classes; see Table 1). For the passive, most errors involved the production of /dɑːn/ (class II) instead of /tɑːn/ or /lɑːn/ (class III).

Statistical analysis of input-based predictors. Here we report the findings of statistical analyses of the simulation data, built to investigate the extent to which the computational model simulates the following findings from the elicited-production studies:

- Effects of both token frequency and phonological neighbourhood density are robust, but there is no evidence for an interaction between these factors.
- Although children's performance increases with age (as would be predicted under any account), there is weak and no evidence respectively that effects of token frequency and phonological neighbourhood density decrease with age.

A regression analysis was performed on a time range from training trial 60,000, where the model reached 75% accuracy overall, until trial 350,000, where all suffixes were produced correctly and only stem errors remained. This range was chosen such that the model — like children — showed relatively low error rates in general while retaining some variability in the error patterns across the system. The epoch range included data from ten separate test points during training from 10 networks each. The mean error rate on the items in the selected range was 9.5%, which is slightly higher than the rate in the experimental data (7.3%).

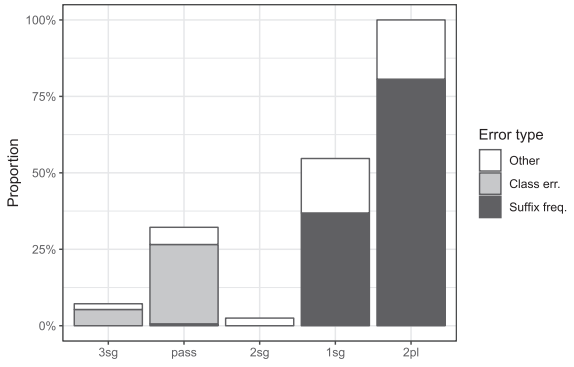

Fig. 6. Finnish model: Proportion of suffix error types per person/number target context for test items after 1000 training trials (suffix mean accuracy at 75%, 32 verbs per P/N, averaged across ten networks). X-axis is arranged by P/N corpus frequency from high to low.

Table 13
Finnish model: Person/number substitution errors in 160 test items after 1000 training trials (suffix mean accuracy at 75%).

	Output			
Target	1sg	2sg	3sg	Pass
1sg		11	106	1
2sg	1		0	0
3sg	3	0		0
Pass	0	0	2	
2pl	72	21	0	165

Table 14
Finnish model: Conjugation class (overgeneralisation) errors in 160 test items after 1000 training trials (suffix mean accuracy at 75%). Only unambiguous errors counted, i.e., where the output ending was assignable to exactly one class.

	Output			
Target	I	II	III	IV
I		11	0	0
II	0		0	0
III	0	14		0
IV	0	8	4	

As in the experiments' analysis, we fitted a logistic mixed-effects model on a binary response variable (correct = 1, incorrect = 0) for the full form, testing the effects of token frequency, class size (PND) and "age" (i.e., number of training trials) as well as all two-way interactions. All predictor variables were log-transformed, scaled and centred. The statistical models included random intercepts for network and verb (lemma). Unlike in the analysis of the empirical data, the use of additional, complex Bayesian models was not necessary here, since less random between-subject error has to be accounted for in simulated data than in human participants. Table 15 reports the frequentist regression results separately for the 160 items the children were tested on (test set) and for all 1784 items in the training corpus (training set).

In the test set, main effects of token frequency and age (training trial) were found but no effect of phonological neighbourhood density (class size). Unlike, however, in the elicited production study, where no significant interactions were observed, all two-way interactions were significant. The model showed a decreasing effect of PND (class size) with increasing token frequency, which is in line with our predictions but was not seen in the experimental study. The modelling results also indicated that the magnitude of the effects of both token frequency and PND increased with training, whereas the experimental study had shown — at least in the pooled analysis — a trend for a decreasing effect of token frequency with age. The interaction effects with training trial in the model are, hence, neither mirroring the data nor in line with our predictions. We look further into these discrepancies in Section 4. The results for the full training set were similar to the test set results, while additionally showing a significant main effect of class size. This suggests that the effect of PND may be too subtle to be detected as a main effect when only considering test items. PND did have some effect, however, even in the test items, as evidenced by the significant interaction of class size with token frequency. The conditional coefficients of all two-way interactions in the training set analysis are plotted in Fig. 7.

In summary, the computational model built for Finnish simulated all of the error patterns and main effects observed in the corresponding child elicited production study, but also yielded additional interactions of token frequency, PND and age that will require further investigation.

Table 15
Regression results for Finnish model accuracy.

	Test items				Training items			
	Est.	SE	$\chi^2(1)$	p	Est.	SE	$\chi^2(1)$	p
(Intercept)	5.24	0.58			6.59	0.16		
ClassSize	0.54	0.58	2.5	0.114	**0.47**	**0.13**	**27.4**	**< 0.001**
TokenFreq	**3.77**	**0.21**	**346.3**	**< 0.001**	**2.19**	**0.09**	**821.1**	**< 0.001**
Trial	**6.27**	**0.72**	**516.0**	**< 0.001**	**7.49**	**0.28**	**2324.2**	**< 0.001**
ClassSize:TokenFreq	**−0.56**	**0.25**	**5.1**	**0.024**	**−0.49**	**0.08**	**41.8**	**< 0.001**
TokenFreq:Trial	**2.19**	**1.00**	**4.8**	**0.028**	**1.77**	**0.37**	**22.7**	**< 0.001**
ClassSize:Trial	**1.47**	**0.26**	**32.9**	**< 0.001**	**0.43**	**0.12**	**12.4**	**< 0.001**

Note: Logistic linear mixed-effects regression results on 160 test and 1784 training items across ten networks at training trials 60,000 – 350,000.

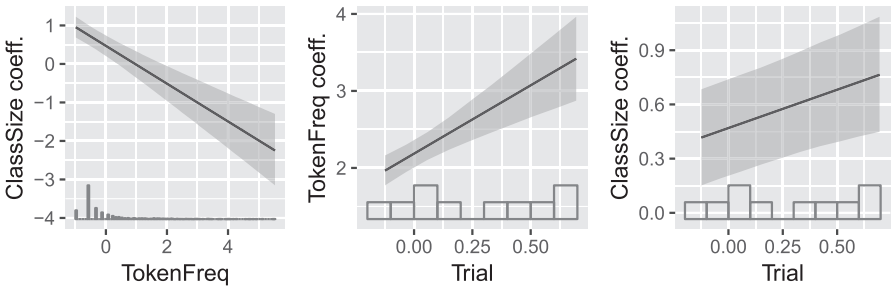

Fig. 7. Conditional coefficients for two-way interaction terms in linear regression on Finnish model accuracy in training set (Table 15). Plotted is the *change of the coefficient* of one predictor (y-axis) as a function of the value of the second predictor (x-axis).

3.5.2. Polish

General performance. Similarly to the Finnish model, all 2431 forms in the Polish training corpus were produced 100% correctly after three million training trials (see Fig. 8). Unlike the Finnish model, the Polish model did not learn suffixes faster than stems. By comparison with Finnish, Polish inflectional suffixes are very irregular, while there are only a few cases where the stem of a verb alternates between inflections.

Generalisation. The Polish model could correctly generalise to 86% of unseen test items at the end of training (4 million trials). Most errors involved not making the correct stem change (e.g., from /vjɛʐ/ to /vjɔzɛ̃/), producing the verb in a more frequent person/number context (e.g., /maʧa/, 3sg, instead of /maʧam/, 1sg), or using a more frequent suffix from a different class (e.g., /maʧɔ̃w/, 1sg-I/II, instead of /maʧajɔ̃w/, 1sg-III). Accuracy on stems and suffixes for new versus trained test items is plotted for seven subclasses in Fig. 9. The lowest rate of stem generalisation (below 75%) was seen in class I-11, a small class with stem alternations. Errors in this class mostly consisted of missing or insufficient stem change. The generalisation of suffixes was poorest in the smallest class IV and the largest class III. In class IV, the model erroneously used suffixes from a larger class for rare person/number contexts, such as 2pl and 3pl. Since class IV is so small, the model had hardly any data for such rare contexts which it could use as a basis for generalisation. In class III, some rarer person/number contexts tended to be produced with suffixes from a more frequent context or with a more frequent suffix from another class.

Descriptive analysis of errors. As for Finnish, we present a descriptive analysis of errors designed to investigate whether two findings of the elicited production studies are simulated by the model:

- Overall error rates are low, but are high for rare person/number contexts (see Section 1.1.1).
- Most errors are of one of three types: (a) frequency-based substitutions; (b) near-miss errors; (c) conjugation-class errors (see Section 1.1.3).

As for Finnish, the Polish model showed a very similar pattern to Polish children with regard to the patterning of error rates by frequent versus infrequent person/number contexts (see Fig. 10). Again, high frequency 3sg and 2sg forms are acquired fastest, while low frequency plural suffixes are learned more slowly, with 2nd plural being particularly hard to learn. Also echoing a finding observed for Finnish, Polish children's over-performance for 1sg forms (relative to their frequency in the input) was not simulated by the model. Again, this is presumably because children are more interested and therefore more highly practised in talking about themselves and perhaps more sensitive to 1sg forms in the input. Unlike in Finnish, however, Polish children also over-performed (relative to their frequency in the input) on 1pl forms. A similar pragmatic explanation may apply here, as such forms also allow children to talk about themselves (albeit themselves plus another person).

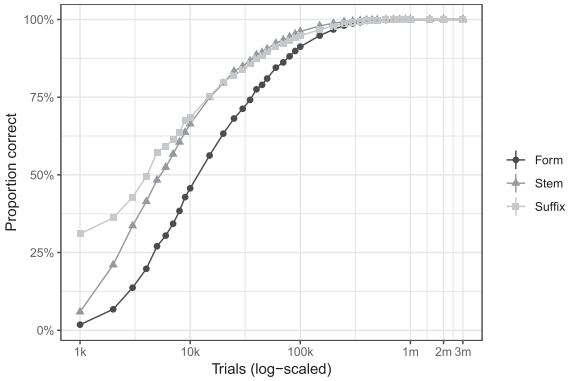

Fig. 8. Polish model: Proportion of correctly produced stems, suffixes and full forms over training (means and standard errors, averaged across ten runs).

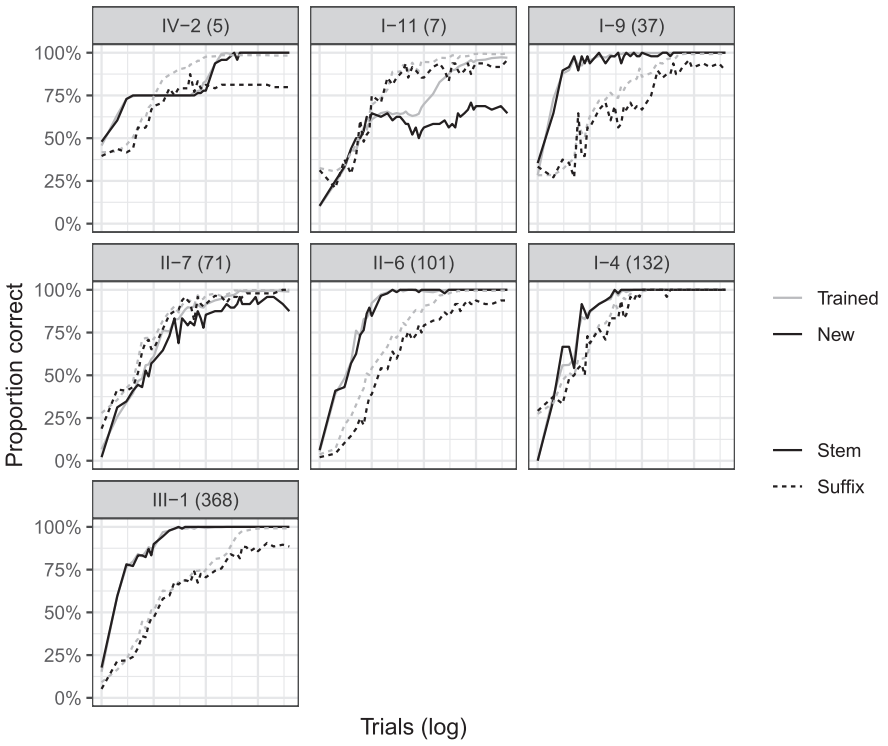

Fig. 9. Polish model: Stem and suffix accuracy on 192 test items when they were part of the corpus (trained) vs. when they were withheld from the corpus (new) items (averaged across ten runs; numbers in parentheses are inflectional neighbourhood sizes with respect to the training corpus).

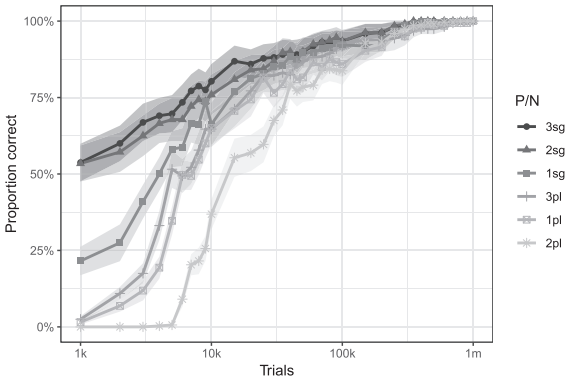

Fig. 10. Polish model: Development of suffix accuracy in test items by person/number target context over training (means and standard errors, averaged across ten networks). Legend ordering (top to bottom) and line shading (dark to light) represent P/N corpus frequency from high to low.

As for Finnish, we conducted a descriptive analysis of error types by interrogating the model at the point at which the mean suffix accuracy was 75% (20,000 training trials), as shown in Fig. 11 and Tables 16 and 17.

The Polish model — like Polish children — differed from Finnish in that conjugation class errors were more common than frequency-based substitutions (c.f., Tables 5 vs. 6). This difference is not unexpected, given that, in Polish, such errors are possible, in principle, for all person/number contexts (see Table 2), whereas, in Finnish, they are possible only for 3sg contexts (and passive contexts, which were not used as targets in the elicitation study). As for the Polish children, most of the Polish model's conjugation class errors (see Table 17) involved the production of class III forms in place of class I targets. These substitutions entail replacing the first suffix vowel /ɛ/ with /a/, as in /aʃ/ instead of /ɛʃ/. The model additionally produced conjugation class errors that were not produced by children: the use of class II suffixes in place of class I and class III.

Frequency-based substitutions (see Table 16) were less common than for Finnish, and all involved the production of a 3sg or 1pl target in place of a 1sg or 2pl form (thus, all but two are also analysable as near-miss errors). Although these errors were also produced by Polish children, the model did not — at least not in this stage of training — simulate another common error produced by the children: the production of 3pl forms in 2pl contexts.

Statistical analysis of input-based predictors. As for Finnish, we ran a linear regression on the computational model's output in order

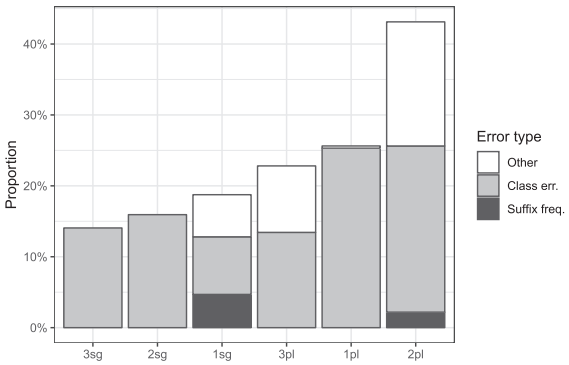

Fig. 11. Polish model: Proportion of suffix error types per person/number target context for test items after 20,000 training trials (suffix mean accuracy at 75%, 32 verbs per P/N, averaged across ten networks). X-axis is arranged by P/N corpus frequency from high to low.

Table 16
Polish model: Person/number substitution errors in 192 test items after 20,000 training trials (suffix mean accuracy at 75%).

	Output	
Target	3sg	1pl
1sg	15	0
2pl	2	5

Table 17
Polish model: Conjugation class (overgeneralisation) errors in 192 test items after 20,000 training trials (suffix mean accuracy at 75%). Only unambiguous errors counted, i.e., where the output ending was assignable to exactly one class.

	Output			
Target	I	II	III	IV
I		61	47	1
II	0		4	1
III	0	33		0
IV	0	1	1	

to investigate the extent to which the model simulates the following findings from the elicited-production studies:

- Effects of both token frequency and phonological neighbourhood density are robust, but there is no evidence for an interaction between these factors.
- Although children's performance increases with age (as would be predicted under any account), there is weak and no evidence, respectively, that effects of token frequency and phonological neighbourhood density decrease with age.

In order to match the Finnish analysis, we obtained model data from a range spanning 75% overall accuracy to 100% suffix accuracy, though — due to the less regular nature of Polish — this corresponded to more training trials (40,000 – 1,500,000), including 22 separate time points of ten networks each. The mean error rate on the test items in the selected range was 8.7%, which is close to the rate in the Polish empirical data (7.8%). Again, the dependent variable was correct (1) vs. incorrect (0) with respect to the target form, and the independent variables were token frequency, phonological neighbourhood density (class size), "age" (training trial), and all two-way interactions. Analyses were performed separately for the 192 test items and for on all forms in the training corpus with available subclass information (2168 out of a total of 2431 training items); see Table 18 for results.

As for Finnish, effects of token frequency and age (trial) were observed in both the test set and the training set, while a main effect of phonological neighbourhood density (class size) was found in the training set only. All two-way interactions were significant in both sets. The conditional coefficients of the interactions in the training set analysis are plotted in Fig. 12. Unlike in the Finnish model, however, the interaction between token frequency and class size was positive, indicating that the effect of phonological neighbourhood density was stronger for highly frequent forms. This does not conform to our predictions, but mirrors the empirical data to the extent that this interaction was positive there, too, although not significant (but with a probability of 0.83 in the pooled Bayesian analysis). Unlike in the Finnish model, the interaction between class size and training trial was negative in the Polish test set, suggesting that the PND effect decreases with training. This conforms to our predictions. However, the analysis of the Polish

Table 18
Regression results for Polish model accuracy.

	Test items				Training items			
	Est.	SE	$\chi^2(1)$	p	Est.	SE	$\chi^2(1)$	p
(Intercept)	3.41	0.38			4.03	0.10		
ClassSize	−0.06	0.39	0.0	0.967	**0.31**	**0.11**	**4.9**	**0.026**
TokenFreq	**1.20**	**0.05**	**473.2**	**< 0.001**	**1.38**	**0.02**	**4086.2**	**< 0.001**
Trial	**4.52**	**0.14**	**2769.0**	**< 0.001**	**5.95**	**0.05**	**22683.3**	**< 0.001**
ClassSize:TokenFreq	**0.32**	**0.05**	**36.7**	**< 0.001**	**0.09**	**0.01**	**47.1**	**< 0.001**
TokenFreq:Trial	**0.91**	**0.16**	**32.2**	**< 0.001**	**0.88**	**0.07**	**165.8**	**< 0.001**
ClassSize:Trial	**−0.49**	**0.08**	**33.1**	**< 0.001**	**0.45**	**0.02**	**354.2**	**< 0.001**

Note: Logistic linear mixed-effects regression results on 192 test and 2431 training items across ten networks at training trials 40,000 – 1,500,000.

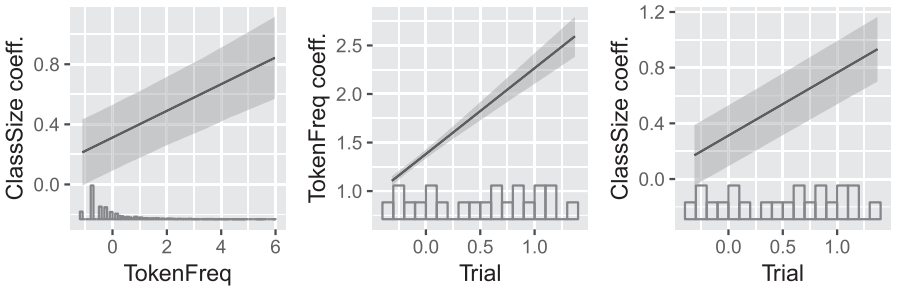

Fig. 12. Conditional coefficients for two-way interaction terms in linear regression on Polish model accuracy in training set (Table 18). Plotted is the *change of the coefficient* of one predictor (y-axis) as a function of the value of the second predictor (x-axis).

training set showed the interaction to be positive like in Finnish.

These findings suggest that either our predictions about interactions between token frequency, PND and age were wrong for Polish, or the observed effects in both the elicited production data and the model output may be misleading due to the low overall error rate and the use of a binary measure of error. In addition, the absence of a main effect of PND in the test set and the finding of a positive interaction of PND with token frequency may be due to discrepancies in type frequencies between the training corpus and the reference grammar, such that the grammar-based class size predictor is not representative of the actual input to the model. These questions will be addressed in Section 4.

3.6. Computational modelling summary

In the Introduction, we outlined three criteria that must be met by any successful model of the acquisition of person/number marking inflectional morphology. The first is that the model must eventually achieve essentially perfect mastery of the system, since native-speaking adults never (or only extremely rarely) produce incorrectly inflected verb forms. Both the Finnish and Polish models met this criterion, achieving 100% accuracy by the end of training. The second criterion is that the model must generalise to unseen items. Again, both the Finnish and Polish models met this criterion with about 85% accuracy by the end of training.

The third criterion is that the model must simulate a number of empirical phenomena observed in studies of children's acquisition of inflection. On the basis of two large-scale elicited production studies, we established that these phenomena are as follows:

- Overall error rates are low, but are high for rare person/number contexts.
- Most errors constitute (a) frequency-based substitutions, (b) near-miss errors, or (c) conjugation-class errors.
- Effects of both token frequency and phonological neighbourhood density (PND) are robust, but there is no evidence for an interaction between these factors.
- Although children's performance increases with age (as would be predicted under any account), there is weak and no evidence respectively that effects of token frequency and phonological neighbourhood density decrease with age.

In general, the Finnish and Polish models simulated all of these phenomena, but also provided some evidence to suggest that the effect of PND may decrease with increasing token frequency in Finnish, but increase in Polish. The former was not seen in the data but conforms to our theoretical predictions, while the latter mirrors a trend in the data but is at odds with our predictions. The simulations also suggested that the effects of both PND and token frequency may — except in the Polish test set — increase with age, which again is at odds with our predictions and partly with the experimental results. We suspect that these discrepancies may be due to an insufficient sample size or the use of an insufficiently sensitive binary dependent measure (correct/incorrect) in both the experiments and the modelling, and we further investigate this in the next section. Another discrepancy between model and data was

that an effect of PND was found in the Polish experimental data but not in the Polish modelling results when analysing just the test items. This is probably due to differences in type frequencies (and therefore class sizes) between the training corpus and the reference grammar. Finally, the model was unable to simulate a phenomenon whereby children showed significantly better performance with 1sg forms (and, for Polish only, 1pl forms) than one would expect given their input frequency. This finding most likely has a pragmatic explanation: children prefer to talk about themselves rather than other people, and so have a great deal of practice in using — or input sensitivity for — 1sg (and perhaps also 1pl) forms, as well as a greater motivation to learn them.

In the following section, we present an alternative analysis of the models' performance and a more detailed analysis of how the model yielded the observed effects.

4. Further investigations of the modelling results

Computational modelling is most useful, in our view, when it does not end at the point at which some empirical data has been simulated, but rather loops back to suggest future potential modifications to our theories that should be tested in empirical data. In other words, we should ask not only "what do children do that the model also does?", but "what does the model do that children might also do?".

Some differences between the computational model and the elicitation experiments have been observed above, i.e., the negative interaction between token frequency and PND in Finnish and the positive interactions of both token frequency and PND with age (training trial) in both languages that were observed in the modelling but not the child empirical data. While these discrepancies may reflect differences in the way the child and the model are learning, it is also possible that they reflect differences in the available data, and that a similar pattern of interactions would emerge in experiments if more children or a greater number of verbs were tested. A general concern applying to both the experiments and the modelling results is that some effects may be misleading due to the low overall error rate and the use of a binary correct/incorrect dependent measure, which can lead to ceiling effects. Another concern is that the class size values of the grammar-based PND predictor may not be representative of the distribution in the training corpus. In what follows, we therefore analysed the modelling output more thoroughly by using a more sensitive, continuous measure of accuracy instead of a binary one, and replacing the grammar-based PND predictor with a corpus-based measure.

4.1. Using more sensitive measures

We performed the same linear regression analysis on the computational modelling data for Finnish and Polish as above. However, instead of a binary measure for correctness of the full form, we used the model's error in terms of the difference between output and target activation, i.e., the cross-entropy error. The error measure is more sensitive to variation in the model's performance, as it captures not only the binary distinction between correct and incorrect but also the degree of deviation from the target in both the correct and incorrect outputs. In order to allow for the same interpretation of the direction of an effect as in the previous analyses, we defined the dependent variable as the (scaled and centred) negative logarithm of the cross-entropy error. This results in a continuous measure of accuracy: The higher the value the closer the output activation pattern is to the target.

The regression results with a continuous response measure are presented in Appendix Tables C.29 and C.30. The Finnish results were similar to the results with the binary measure, except that now the effect of token frequency decreased with more training in both the test set and the training set; and the effect of PND decreased over training in just the training set. Similarly, the Polish results with a continuous response measure now showed decreasing effects of token frequency and PND in both the test set and the training set. These results show that, in both languages, the interactions of token frequency and PND with age (training trial) tend to be in the predicted direction when using a continuous response measure to analyse the simulation data. However, in Polish, the interaction between token frequency and PND remains positive (against predictions) and an effect of PND was not found at all when using the continuous response measure.

As we stated earlier, a PND predictor based on type frequencies from a reference grammar may be inaccurate for the modelling data. Because the child-directed speech corpora in Finnish and Polish represent a particular subset of spoken language, type frequencies are expected to be different from the reference grammars. Class sizes based on the grammars may still be a good predictor for children's production data, since children obviously receive much more input than what is recorded in a CDS corpus. The model, however, is restricted to exactly the input provided by the training corpus. Therefore, major differences in the class ranks could dramatically reduce the correlation of the grammar-based PND predictor with model accuracy. As the class ranks in Tables 19 and 20 show, there are some discrepancies between the type frequencies in the grammars and the corpora in both Finnish and Polish. Also, the fact that Polish has only seven subclasses versus eleven in Finnish may add to class size being a weaker predictor in Polish than in Finnish. We therefore performed regression analyses on the modelling data using binary and continuous responses with a corpus-based class size measure instead of a grammar-based measure.

Results with corpus-based PND and a *binary* response measure are provided in the Appendix for completeness (Tables D.31 and D.32). These were similar to the binary response analyses with a grammar-based class size measure, except that the main effect of class size in Finnish was now significant in both the training set and the test set. The results with the most sensitive measures — corpus-based class size and a *continuous* response — are presented in Tables 21 and 22. In the training sets of both languages, all main effects were found and all interactions were now negative (see conditional coefficients in Fig. 13), indicating that the effect of PND decreased with both increasing token frequency and increasing age; the effect of token frequency also decreased with increasing age. In the Polish test items, however, we still found main effect of PND, and the interaction of PND and token frequency was still positive. PND effects in the Polish model, therefore, may be too subtle to be detected in just the test items, at least when measured by the type

Table 19
Finnish inflectional subclasses and their ranks based on verb type frequency (in parentheses) in the reference grammar and the training corpus.

Subclass	Grammar	Corpus
I-d	1 (3093)	1 (230)
I-a	2 (2228)	2 (163)
III-a	3 (1329)	3 (108)
IV-c	4 (885)	4 (92)
II-b	5 (730)	9 (9)
I-b	6 (402)	5 (60)
III-b	7 (272)	6 (59)
IV-b + d	8 (170)	7 (31)
IV-e	9 (49)	11 (5)
I-c	10 (31)	8 (19)
II-a	11 (15)	10 (9)

Table 20
Polish inflectional subclasses and their ranks based on verb type frequency (in parentheses) in the reference grammar and the training corpus.

Subclass	Grammar	Corpus
I-4	1 (6499)	2 (132)
III-1	2 (6489)	1 (368)
II-6	3 (5767)	3 (101)
I-9	4 (1473)	5 (37)
I-11	5 (1261)	6 (7)
II-7	6 (401)	4 (71)
IV-2	7 (10)	7 (5)

Table 21
Finnish model continuous accuracy with corpus-based class size measure.

	Test items				Training items			
	Est.	SE	$\chi^2(1)$	p	Est.	SE	$\chi^2(1)$	p
(Intercept)	0.28	0.085			0.43	0.010		
ClassSize	**0.25**	**0.085**	**6.2**	**0.013**	**0.13**	**0.011**	**77.9**	**< 0.001**
TokenFreq	**0.23**	**0.006**	**772.2**	**< 0.001**	**0.13**	**0.002**	**2936.6**	**< 0.001**
Trial	**0.51**	**0.011**	**2255.4**	**< 0.001**	**0.36**	**0.003**	**19660.8**	**< 0.001**
ClassSize:TokenFreq	**−0.05**	**0.005**	**120.7**	**< 0.001**	**−0.03**	**0.001**	**699.0**	**< 0.001**
TokenFreq:Trial	**−0.31**	**0.011**	**862.1**	**< 0.001**	**−0.23**	**0.003**	**7745.3**	**< 0.001**
ClassSize:Trial	**−0.13**	**0.011**	**157.1**	**< 0.001**	**−0.12**	**0.003**	**2219.1**	**< 0.001**

Note: Linear mixed-effects regression results on 160 test and 1784 training items across ten networks at training trials 60,000 – 350,000.

Table 22
Polish model continuous accuracy with corpus-based class size measure.

	Test items				Training items			
	Est.	SE	$\chi^2(1)$	p	Est.	SE	$\chi^2(1)$	p
(Intercept)	0.225	0.051			0.287	0.010		
ClassSize	0.135	0.051	2.2	0.14	**0.126**	**0.010**	**51.0**	**< 0.001**
TokenFreq	**0.263**	**0.004**	**998.1**	**< 0.001**	**0.202**	**0.001**	**8867.0**	**< 0.001**
Trial	**0.464**	**0.004**	**10844.6**	**< 0.001**	**0.478**	**0.001**	**169475.1**	**< 0.001**
ClassSize:TokenFreq	**0.014**	**0.003**	**25.1**	**< 0.001**	**−0.005**	**0.001**	**52.6**	**< 0.001**
TokenFreq:Trial	**−0.315**	**0.004**	**4953.4**	**< 0.001**	**−0.243**	**0.001**	**43097.6**	**< 0.001**
ClassSize:Trial	**−0.122**	**0.004**	**739.3**	**< 0.001**	**−0.104**	**0.001**	**7867.2**	**< 0.001**

Note: Linear mixed-effects regression results on 192 test and 2431 training items across ten networks at training trials 40,000 – 1,500,000.

count of an inflectional class.

Except for the Polish test items, all main effects and interactions are in line with our predictions. It is possible that important error variance was being missed when analysing the computational model using a binary measure and a grammar-based measure of PND, and some of the effects were therefore misleading, in particular, the positive interaction between token frequency and PND in Polish

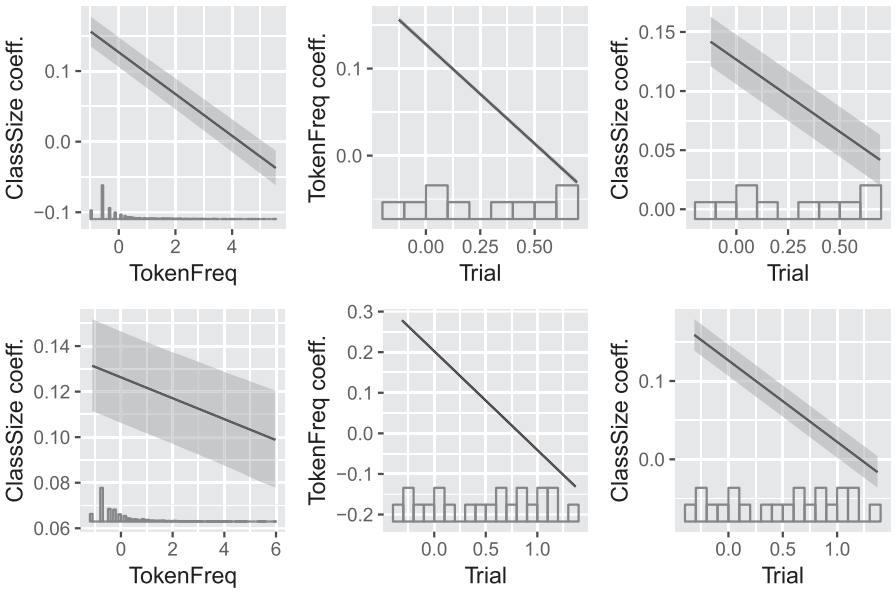

Fig. 13. Conditional coefficients for two-way interaction terms in linear regression with corpus-based class sizes on Finnish (top row) and Polish (bottom row) model continuous accuracy in training set (Tables 21 & 22). Plotted is the *change of the coefficient* of one predictor (y-axis) as a function of the value of the second predictor (x-axis).

and the positive interactions of PND or token frequency with training trial in both languages. It remains to be shown whether interaction effects in the elicited-production studies can also be observed with a larger sample size or a more fine-grained measure of error than simply correct/incorrect.

4.2. Hidden layer activation patterns

In order to better understand how the model achieves phonologically-based generalisation, we analysed the activation patterns of the hidden layer. These activation patterns are a representation of the abstracted knowledge the network uses to produce — when trained to criterion — the correct form. This representation is abstract in the sense that it does not encode all information presented on the input and output layers during training, but only the features the network has discovered to be relevant for the mapping task. During training, the connection weights become tuned such that verbs that have a similar mapping from input to inflected form are represented by a similar pattern in the hidden layer. In other words, the representations in the model's hidden layer correspond to the kind of inflectional "rules" posited under formal symbolic approaches (or described in reference grammars), albeit in a less categorical way.

We performed a hierarchical clustering analysis on the activations of the 200 hidden units of one randomly chosen network in each language after training for 4 million trials. We first calculated the Euclidean distances between the hidden unit activations of all test verb inputs and then, based on these distances, hierarchically divided the input items into clusters using the R function `hclust` (R Core Team, 2016) with Ward's minimum variance method. The dendrograms in Fig. 14 show the hierarchical clustering of the 32 test verbs for Finnish and Polish. The figure shows 3sg forms in Finnish and 1sg forms in Polish, because class membership is specifically relevant in these contexts: inflection suffixes in Finnish differ between classes only for 3sg and the passive; and, in Polish, stem alternations occur only in 1sg and 2sg (in subclasses I-11 and II-7).

It is clear from Fig. 14 that overlap between the clustering and inflectional subclasses (as of the reference grammar used to classify verbs into phonological neighbourhoods for the elicited production study) is more obvious in Polish. The reason for seeing less class/cluster overlap in Finnish may be that (a) Finnish person/number marking suffixes are much less dependent on inflection class than in Polish and (b) Finnish stem changes are — due to consonant gradation — less predictable than in Polish.

There is one class in particular in Polish — class IV — which is perfectly represented, meaning that there is a cluster that includes all and exclusively class IV instances. This makes the class relatively easy to learn (see Fig. 9) despite the fact that it is the smallest class. The grouping of class IV items in Polish is facilitated by the fact that all verb stems in this class are short and end in the same consonant /j/. Class IV is grouped closest to class I-4, all stems of which end in /uj/. From this and elsewhere in Fig. 14 it is apparent that the model — especially in Polish — grouped verbs mainly on the basis of similarities in the stem *rhyme*, often across subclass borders, such as /ɛʒ/ – /ɛntɕ͡ / – /aɲt͡ʃ / – /at͡ʃ / – /aʒ/. These examples all end in a postalveolar or alveolo-palatal fricative but are from different subclasses (I-11, II-7, II-6, III-1 and I-9). The inflectional subclasses in reference grammars are based on the stem rhyme, too, but take into account additional factors that might not be available to the model or picked up by it (such as stem alternations within and across tenses). The clusterings across class borders mostly make sense in the context of the model's task (most of the above examples build their 1sg with the suffix /-ɛ̃/) but sometimes include verbs with dissimilar suffixes (e.g., /mat͡ʃ-am/), so-

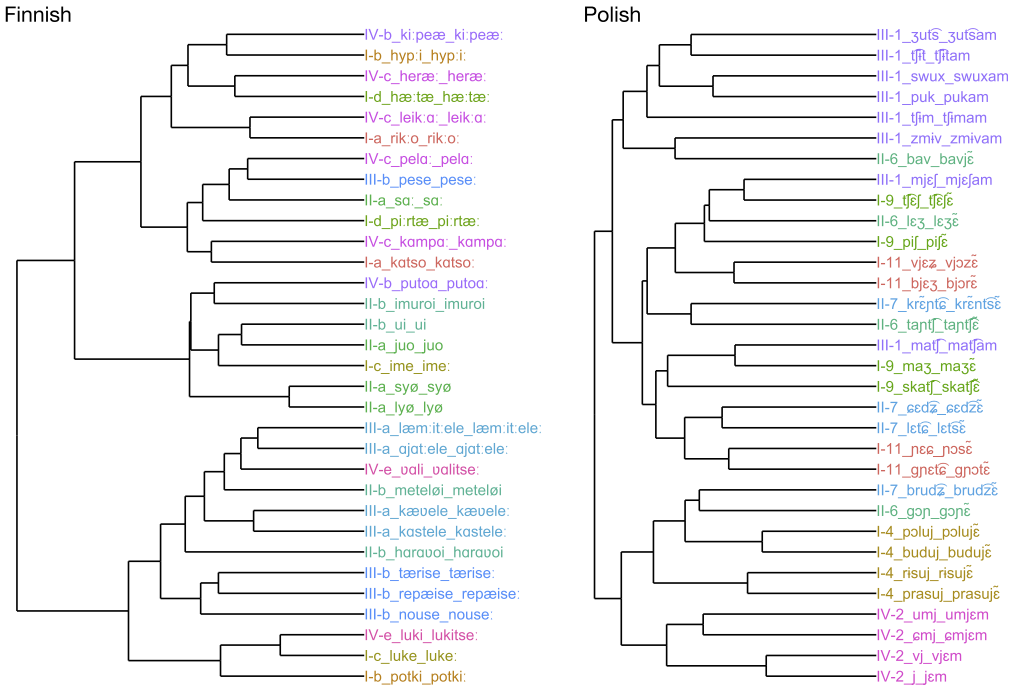

Fig. 14. Hierarchical clustering of the hidden layer representations for 32 test verbs in Finnish (left) and Polish (right). Verbs are annotated with subclass, input stem and the inflected form (FI: 3sg; PL: 1sg). Colours highlight subclass membership. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

called “enemies”, (e.g., Marchman, 1997; Marchman et al., 1999). The neighbourhood density measure used in the experiments and modelling here and in previous studies (e.g., Mirković et al., 2011; Räsänen et al., 2016; Savičiūtė et al., 2018) — i.e., the number of verbs per inflection class — ignores possible enemies in the neighbourhood. However, the number of phonological enemies in terms of the stem rhyme has been shown to affect children’s productivity in the English past tense (Marchman, 1997; Marchman et al., 1999). Our expectation is therefore that a verb that is surrounded by enemies would be inflected with less accuracy by the model. For example, the Polish 1sg form /maʧam/, which is surrounded by enemies in Fig. 14, is amongst the forms that could not be successfully generalised.

It is important to note that the internal representations are influenced not only by the similarity structure in the corpus but also by the way the input and output are presented to the model. Thus, the clustering may be different when stems are not represented in the input or when the phoneme templates are not right-justified. However, as explained in the modelling section, there are particular motivations behind all of our architectural choices. The clustering, therefore, represents the similarity structure of the input as constrained by a learning model with a certain degree of psychological plausibility. As such, the results of the clustering suggest that, by focusing only on class membership, the PND measures used in the current study may fail to capture some of the phonological structure in the input that drives the model’s and, by hypothesis, the child’s performance, and that future work should therefore use more sophisticated measures of PND when investigating the role of PND in predicting the pattern of error in children’s speech. To devise such a measure for large corpora of complex inflection paradigms like the ones we have studied here is more complicated than for the English past tense and, hence, involves the application of sophisticated computational methods. We demonstrate the feasibility of such a fine-grained, computational measure elsewhere in an elicited-production study of noun case marking in Finnish, Polish and Estonian in Granlund et al. (submitted for publication).

4.3. Frequency versus regularity

Here, we demonstrate the effect of suffix regularity in Finnish by separating it from frequency effects through manipulation of the training corpus. As noted above, the passive (1pl) in Finnish is learned relatively slowly by the model despite being the second most frequent person/number context after 3sg in the training corpus; a finding that we provisionally attributed to its irregularity: possible passive endings are /tɑːn/, /dɑːn/ and /lɑːn/. On the other hand, 3sg is learned relatively quickly, even though it is also irregular: the final stem vowel is lengthened unless the stem ends in a diphthong or long vowel. All other person/number forms are regular, i.e., use the same ending across classes. In order to investigate a possible frequency/regularity trade-off, we ran additional simulations where we removed differences in token frequency and class size from the input corpus. We constructed a balanced training corpus where the number of tokens by conjugation class and by person/number combination was equal. The corpus contained eleven forms per person/number combination in each of the four classes (i.e., 220 forms in total). In training, every form was presented to the networks with the same frequency.

Trained on the balanced corpus, the model passed 99% accuracy after only 10,000 trials, but needed until 250,000 trials to reach 100% accuracy. This suggests that there are a few unusually complicated forms that needed more than twenty times more exposure to be mastered than the majority. In the absence of differences in token frequency or neighbourhood effects, the person/number targets acquired fastest were now 1sg and 2pl (as opposed to 3sg), while the passive (1pl) remained the slowest. This analysis demonstrates how the interaction of frequency and regularity effects on acquisition can be disentangled and suggests that the variability of suffixes within a person/number context (or other morphological context) is a measure to be considered when predicting children's production errors.

4.4. Exploring the effect of a localist (vs. distributed) input representation

As we noted in Section 3.1, the decision to map from stem to inflected form, which was taken for mainly implementational reasons, does not entail the assumption that children start out with a bare stem form to which they subsequently apply an inflection. We argued that this distributed input representation is justified on the basis that the lexical semantics of a verb (e.g., sleeping) cue all of the various phonological forms that contain the stem (e.g., Finnish *nukun*, *nukut*, *nuku* etc.). The alternative — a localist representation, in which each verb is represented by a different input unit (e.g., as in the noun case marking study of Mirković et al., 2011) — assumes that learners have no knowledge of the phonological properties of a lemma, independent of each particular person/number form.

In order to explore whether our findings depend crucially on the use of a distributed input representation, we ran additional simulations using a localist representation. In general, the findings were very similar (see Appendix E for detailed results). Importantly, all five major findings were replicated: (1) Eventual perfect acquisition of the system; (2) Good generalisation to unseen forms of previously-encountered verbs; (3) Error rates low overall, but high for some person/number contexts; (4) Most errors reflect (a) frequency-based substitutions; (b) near-miss errors; (c) conjugation-class errors; (5) Effects of age, token frequency and phonological neighbourhood density. The details of the localist model and the simulation results can be found in Appendix E.

Of course, a major advantage of having phonological information on the input layer in the distributed-input model is that it can (at least in principle) succeed on a *Wug test*, i.e., generalise to unknown verbs, which the localist model cannot. In the present study, we are only testing novel forms of *known* verbs in both models. We leave it to a future study to gather empirical data on children's Wug test capability in Finnish and Polish and compare it with the model.

5. General discussion

The aim of the present work was to build a computational model of children's acquisition of present tense person/number marking in two highly inflected languages (Finnish and Polish). First, in order to establish which putative learning phenomena are sufficiently robust to constitute a target for modelling, we ran studies in which native learners of Finnish ($N = 77$; 35–63 months) and Polish ($N = 81$; 35–59 months) were prompted to attempt to produce all five (Finnish) or six (Polish) commonly used person/number forms of each of 32 verbs. We found that

- Overall error rates are low, but are high for rare person + number contexts.
- Most errors constitute (a) frequency-based substitutions, (b) near-miss errors, or (c) conjugation-class errors.
- Effects of both token frequency and phonological neighbourhood density (PND) are robust, but there is no evidence for an interaction between these factors.
- Although children's performance increases with age, there is weak and no evidence respectively that effects of token frequency and phonological neighbourhood density decrease with age.

We then built, for each language, a connectionist model designed to simulate (1) eventual perfect adult-like acquisition of the system, (2) generalisation to novel forms and (3) the developmental phenomena observed in the present elicited production studies. In general, the models — including those implemented using an alternative localist representation — successfully met all three criteria. This work therefore demonstrates that computational approaches largely developed for simpler systems such as English past-tense (Daugherty & Seidenberg, 1992, 1994; Forrester & Plunkett, 1994; Hahn & Nakisa, 2000; Hare & Elman, 1995; Hare et al., 1995; Joanisse & Seidenberg, 1999; Joanisse & McClelland, 2015; MacWhinney & Leinbach, 1991; Plunkett & Bandelow, 2006; Plunkett & Juola, 1999; Plunkett & Marchman, 1991, 1993, 1996; Plunkett & Nakisa, 1997; Plunkett et al., 1992; Ruh & Westermann, 2008; Rumelhart & McClelland, 1985; Westermann & Ruh, 2012) and case (Mirković et al., 2011) can scale-up to the full complexities of person/number marking in highly inflected languages. It also provides evidence to suggest why children learning highly inflected languages show early and extremely accurate marking on highly frequent forms while taking much longer to acquire the full system. This pattern is clearly predicted by the input-based learning architecture used here.

At the same time, two potentially important discrepancies between the present empirical and modelling findings were observed. First, when using a continuous measure of accuracy and a corpus-based measure of PND, both the Finnish and the Polish model provided evidence to suggest that the effect of PND may decline with increasing token frequency, and that the effects of both PND and token frequency may decrease with age. These findings were not observed in the elicited production studies and only partially in the simulations when using a binary dependent measure of correct versus incorrect, suggesting that a binary measure may not be sufficiently sensitive. Future studies (and/or more detailed phoneme-by-phoneme reanalyses of the present child production data) should explore this possibility.

One possible problem with a binary response measure is that, with very low error rates, there may not be enough variability in the responses. Specifically the incorrect responses may have degrees of deviation from the correct response that are not retained by a binary mapping. In the case of the modelling, the binary mapping from an output activation distribution to a particular form — i.e.,

mapping to closest phoneme — is potentially even unrealistic. We currently do not know how speakers choose a specific form when two forms may be equally likely based on their learning experience. Even if the probability is 50.1% versus 49.9% does the speaker necessarily produce the first, as opposed to picking at random, or producing the most frequent, or most recently-used form? In addition to this, the model, unlike the children, produces non-existent phoneme combinations due to a lack of training on the general phonological rules of the language that children certainly receive from outside the domain of verb marking. A continuous response measure that takes into account the deviation from the correct target instead of a binary correct/incorrect mapping may therefore be more informative at least for the simulated data. This illustrates how these seemingly small analysis decisions can make a big difference to the pattern of findings, and it would be misleading of us to say — post hoc — that the one that yields the predicted pattern of results is the "right" one. The best we can say is that there exists a version of the model that yields the predicted pattern of results; and future work should test pre-registered predictions from this version. We presented all versions of our analyses, such that readers can choose for themselves the interpretation they favour.

The second discrepancy between the empirical and modelling work was that the model was unable to simulate a phenomenon whereby children showed significantly better performance with 1sg forms (and, for Polish only, 1pl forms) than one would expect given their input frequency. This seems to be a common finding (see, e.g., Aguado-Orea & Pine, 2015; Theakston et al., 2005), which most likely has a pragmatic explanation: children prefer to talk about themselves rather than other people, and so have a great deal of practice in using 1sg (and perhaps also 1pl) forms, as well as a greater motivation to learn them. One could potentially simulate this phenomenon in future work by adding children's own output to the model input (though this may be difficult to motivate from a theoretical perspective) or by incorporating some measure of the utility of an utterance for a speaker, perhaps as measured by independent adult raters. For example, the 1sg and 1pl forms *I want…* and *we want…* would presumably achieve higher speaker-utility ratings than the 3sg form *he wants…*, as only the former two can be used to request an outcome that is desired by the speaker.

In addition to simulating adult mastery of the system and most — if not quite all — phenomena observed in the child studies, the present models suggested possible adjustments to the definition of phonological neighbourhood that — if confirmed in further experimental studies with children — may need to be incorporated into theoretical and computational models of the acquisition of inflectional morphology. In particular, analysis of the models' internal representations revealed that verbs were grouped on the basis of phonological similarities between stem rhymes, which included so-called *enemies* from a different class. Children's errors may therefore be better predicted when defining phonological neighbourhood in terms of stem-rhyme similarity on the basis of phonological features, and differentiating between friends (neighbours with the same inflection) and enemies (neighbours with a different inflection), as has been done in earlier studies of the English past tense (e.g., Marchman, 1997; Marchman et al., 1999). A computational approach to a fine-grained neighbourhood measure has been applied to complex noun case marking in Granlund et al. (submitted for publication) and could be applied to verb marking in the same way.

In a further investigation of the model, we manipulated the input corpus to remove differences in token frequency and class size, thus revealing the effect of within-person/number context regularity. This showed that 1sg and 2pl contexts, which are the two least frequent, are the easiest forms to learn on average in Finnish when neither frequency nor neighbourhood plays a role, thus suggesting that the regularity — or suffix variance — within a morphological context should also be taken into account for when predicting children's production errors (see Granlund et al., submitted for publication, for an analysis involving the effects of suffix variance in noun marking). More generally, this analysis demonstrates that learning is directly affected by changes in the frequency distribution of the input. The model training as well as the predictors used for analysing the experiments were both based on the assumption of a fixed frequency distribution. However, it is likely that the input frequency of certain forms or contexts varies particularly in the early stages of acquisition. For predicting these early stages of learning complex inflection, detailed corpus studies are necessary, as well as possibly a model with a discontinuous or incremental training regime as used by Rumelhart and McClelland (1985) and Plunkett and Marchman (1993).

Before concluding our discussion, it is important to address head-on the question of whether, in building the present model, we "build in or presuppose surrogates for the linguistic phenomena [we] claim to eschew" (Pinker & Ullman, 2002: 462). First, consider the objection of an anonymous reviewer that the model "starts with innate (or at least pre-existing) categorical – that is, symbolic – knowledge of person and number". It is certainly true that, from the first training input, the model is sensitive to person/number distinctions (implemented by corresponding input units). However, person and number are not solely formal linguistic distinctions, but correspond to observable properties of the world; distinctions to which both human infants and other primate species are sensitive (e.g., Barner, Wood, Hauser, & Carey, 2008; Gallup, 1998). To be sure, the model is "told" that these particular distinctions are relevant for verb inflection, rather than having to learn to pick them out from a wider set. But there is no need to build a further model to demonstrate that irrelevant input from additional input nodes encoding, for example, the speaker's height or eye colour would be rapidly ignored; this is a fundamental property of connectionist networks.

Second, it is true that the model is given, on the input layer, the "stem" of a verb, for which it then generates an inflected form. Again, however, "stem" exists not solely as a formal linguistic notion, but also as an underspecified phonological representation cued by lexical semantics; as, for example, an English-speaking child's knowledge that every word form that she has understood as meaning something like "hit with the foot" sounds something like (though is not always identical to) "kick". Indeed, given the existence of so-called "stem-change" verbs — verbs for which the element conveying the lexical semantics is *not* invariant between different forms — the notion of a formal stem would seem to be at something of a disadvantage compared to the notion of an underspecified phonological representation (though, of course, this is an empirical question).

Third, it is true that the input to the model uses right-justified templates, which corresponds to the formal symbolic notion of a suffix. Again, however, it also corresponds to non-symbolic phenomena, such as the recency effect (Gallup, 1998; Glanzer, 1982) and the assumption that children learning suffixing languages will have noticed (not necessarily consciously, of course) that words that exist in different forms to express similar meanings (e.g., *plays, played, playing*) tend to have different endings and invariant beginnings.

Finally, it is true that the model uses feature-based phonological representations that correspond to traditional formal symbolic approaches (e.g., Chomsky & Halle, 1968). Again, however, these representations are not incompatible with other approaches, such as exemplar-based phonology (e.g., Pierrehumbert, 2001), and are adopted merely to abstract away from phenomena that are not relevant to the problem in hand, such as inter-speaker variability (e.g., Hay, 2018) (but which, incidentally, are problematic for formal approaches).

Ultimately, of course, the question of what built-in assumptions and operationalizations are necessary for successful acquisition of the system (as opposed to merely convenient) is an empirical one. A related empirical question for future research is whether a model using a traditional symbol-processing architecture could equal, or even better, the performance of the present connectionist model. Certainly, the present model does not yet achieve adultlike accuracy on generalization to novel items (approximately 85% accuracy), leaving room for potential improvement. In the present study, we intentionally used a very basic architecture, but adjustments such as multiple hidden layers (MacWhinney & Leinbach, 1991), recurrent layers (Mirković et al., 2011) or an incremental training regime (Plunkett & Marchman, 1993) could potentially improve performance. A third question for future research is therefore whether a model with similar architecture to the present one could scale up to the entire verb (or noun) system of a highly-inflected language. Although we believe the present study to be the most comprehensive of its type, it still addresses only a relatively small part of each system. For example, with regard to Finnish, Hakulinen et al. (2004) list more than 250 verb inflections, with the noun system more complex still. Note also that both the verb and the noun system in Finnish are subject to vowel harmony, which was neutralised for the present model. It therefore remains to be seen whether rote storage and phonological analogy remain viable when faced with a system of such complexity and magnitude.

In the meantime, the present findings demonstrate that a model with assumptions that — we would argue — are both minimal and ecologically valid can, in principle, account for the acquisition of even relatively complex systems of person/number marking, on the basis of rote storage and phonological analogy. We suggest, therefore, that it is these types of models that hold the greatest promise of explaining not just the acquisition of verb person/number marking, but also the acquisition of inflectional morphology, and of language in general.

Acknowledgements

This work was supported by the International Centre for Language and Communicative Development (LuCiD). The support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged. We are very grateful for the help of Minna Kirjavainen, who collected, transcribed and curated the Kirjavainen-Max Planck Finnish corpus, and Viljami Venekoski, who created the animated stimuli for the Finnish experiment. This work benefited greatly from extensive discussions with Grzegorz Krajewski and Jeffrey Elman. Finally, we thank the children, schools, nurseries, parents and teachers who made this research possible.

Appendix A. Test items and subclasses

See Table A.23–A.26.

Table A.23
The classification of Finnish verbs into subclasses, along with examples (1st infinitive, strong/weak vowel stem, consonant stem) and phonological neighbourhood density (PND) (from Hakulinen et al., 2004). Capital letters denote variation according to vowel harmony (both back and front vowel variants), and 'VV' indicates a long vowel. The final column shows whether the subclass was included in test items in the current study.

Class	Stem end	1st inf	V stem	C stem	PND	
Class I						
a	-O/U	nukkua	nukku/nuku	–	2228	✓
b	-i	leikkiä	leikki/leiki	–	402	✓
c	-e	lähteä	lähte/lähde	-	31	✓
d	-A	piirtää	piirtä/piirrä	-	3093	✓
Class II						
a	-VV/monosyll.	saada	saa	–	15	✓
b	-Oi/Oitse	lapioida	lapioi	–	730	✓
Class III						
a	-le	tulla	tule	tul	1329	✓
b	-se	nousta	nouse	nous	272	✓
c	-kse	juosta	juokse	juos	3	
d	-ne/re	mennä	mene	men	5	
Class IV						
a	-iA	hävitä	häviä	hävit	12	
b	-OA/UA/eA	haluta	halua	halut	170	✓
c	-AA	liimata	liimaa	liimat	885	✓
d	-itse	lukita	lukitse	lukit	49	✓
e	-ne	paeta	pakene	paet	143	

Table A.24
Finnish verb test items in 1st infinitive form.

Class	1st inf.	Translation
Ia	katsoa	'to look'
Ia	rikkoa	'to break'
Ib	hyppiä	'to jump'
Ib	potkia	'to kick'
Ic	imeä	'to suck'
Ic	lukea	'to read'
Id	häätää	'to shoo'
Id	piirtää	'to draw'
IIa	juoda	'to drink'
IIa	lyödä	'to hit'
IIa	saada	'to get'
IIa	syödä	'to eat'
IIb	haravoida	'to rake'
IIb	imuroida	'to vacuum'
IIb	metelöidä	'to make noise'
IIb	uida	'to swim'
IIIa	ajatella	'to think'
IIIa	kastella	'to water'
IIIa	kävellä	'to walk'
IIIa	lämmitellä	'to warm up'
IIIb	nousta	'to rise'
IIIb	pestä	'to wash'
IIIb	repäistä	'to rip suddenly'
IIIb	täristä	'to shiver'
IVb	kiivetä	'to climb'
IVb	pudota	'to fall'
IVc	herätä	'to wake up'
IVc	kammata	'to comb'
IVc	leikata	'to cut'
IVc	pelata	'to play'
IVe	lukita	'to lock'
IVe	valita	'to choose'

Table A.25
The classification of Polish verbs into subclasses, along with examples (infinitive, 1st sg, 2nd sg) and phonological neighbourhood density (PND). PND for some subclasses was not available. The final column shows whether the subclass was included in test items in the current study. PT stands for past tense.

Class	Inf.	1st sg	2nd sg	PND	
Class I					
4	malować	maluję	malujesz	6499	✓
9	mazać	mażę	mażesz	1743	✓
5	ciągnąć	ciągnę	ciągniesz	X	
3	tanieć	tanieję	taniejesz	X	
8	like 4 but diff stem in PT			2228	✓
10a,b	pić	piję	pijesz	X	
10c	like 5 but diff stem in PT			X	
11	wieźć	wiozę	wieziesz	1261	✓
Class II					
6	robić	robię	robisz	5767	✓
7	widzieć	widzę	widzisz	401	✓
Class III					
1	czytaća	czytam	czytasz	6489	✓
Class IV					
2	umieć	umiem	umiesz	10	✓

Table A.26
Polish verb test items in infinitive form.

Class	Inf.	Translation
I4	rysować	'to draw'
I4	budować	'to build'
I4	prasować	'to iron'
I4	polować	'to hunt'
I9	pisać	'to write'
I9	skakać	'to jump'
I9	czesać	'to comb'
I9	mazać	'to scrawl'
I11	brać	'to take'
I11	nieść	'to carry'
I11	gnieść	'to crumple'
I11	wieźć	'to carry (by car/train)'
II6	bawić	'to play'
II6	leżeć	'to lie down'
II6	tańczyć	'to dance'
II6	gonić	'to chase'
II7	siedzieć	'to sit down'
II7	lecieć	'to fly'
II7	kręcić	'to spin'
II7	brudzić	'to stain'
III	słuchać	'to listen'
III	czytać	'to read'
III	trzymać	'to hold'
III	rzucać	'to throw'
III	pukać	'to knock'
III	mieszać	'to stir'
III	zmywać	'to wash up'
III	maczać	'to dip'
IV	wiedzieć	'to know'
IV	jeść	'to eat'
IV	umieć	'to can/to know how to'
IV	śmieć	'to dare'

Appendix B. Binary phoneme representations

See Tables B.27 and B.28.

Table B.27
Binary phoneme coding for consonants used in both Finnish and Polish models. For lengthened phonemes (/ː/) the last unit is set to 1.

IPA	Binary code							
p	1	0	0	0	0	1	0	0
b	0	0	0	0	0	1	0	0
β	0	0	0	0	1	0	0	0
t	1	0	1	1	0	1	0	0
d	0	0	1	1	0	1	0	0
k	1	1	1	0	0	1	0	0
g	0	1	1	0	0	1	0	0
m	0	0	0	0	0	0	1	0
n	0	0	1	1	0	0	1	0
ɲ	0	1	0	1	0	0	1	0
ŋ	0	1	1	0	0	0	1	0
h	1	1	1	1	1	0	0	0
r	0	0	1	1	1	0	1	0
l	0	0	1	1	1	1	0	0
j	0	1	0	1	0	1	1	0
w	0	0	0	0	0	1	1	0
f	1	0	0	1	1	0	0	0
v	0	0	0	1	1	0	0	0
ʋ	0	0	0	1	0	1	1	0
S	1	0	1	1	1	0	0	0
Z	0	0	1	1	1	0	0	0
ʃ	1	1	0	0	1	0	0	0
ʒ	0	1	0	0	1	0	0	0

(continued on next page)

Table B.27 (*continued*)

IPA	Binary code							
ɕ	1	1	0	1	1	0	0	0
ʑ	0	1	0	1	1	0	0	0
X	1	1	1	0	1	0	0	0
ç	1	1	0	1	1	0	0	0
ð	0	0	1	0	1	0	0	0
ɣ	0	1	1	0	1	0	0	0
ʎ	1	1	0	1	1	1	0	0
ʁ	0	1	1	1	1	0	0	0
θ	1	0	1	0	1	0	0	0
ɥ	1	1	0	1	0	0	0	0
ʔ	1	1	1	1	0	1	0	0
t͡ʃ	1	0	1	1	1	1	1	0
d͡ʒ	0	0	1	1	1	1	1	0
t͡ʃ	1	1	0	0	1	1	1	0
t͡ʒ	0	1	0	0	1	1	1	0
t͡ɕ	1	1	0	1	1	1	1	0
t͡ʑ	0	1	0	1	1	1	1	0

Table B.28
Binary phoneme coding for vowels used in both Finnish and Polish models. For lengthened phonemes (/_:/) the last unit is set to 1.

IPA	Binary code						
i	0	1	0	0	1	0	0
y	0	1	1	0	1	0	0
ɨ	1	1	0	0	1	0	0
ʉ	1	1	1	0	1	0	0
u	1	0	1	0	1	0	0
ɪ	0	1	0	0	1	0	0
ʏ	0	1	1	0	1	0	0
ʊ	1	0	1	0	1	0	0
e	0	1	0	1	1	0	0
ø	0	1	1	1	1	0	0
ə	1	1	0	1	1	0	0
o	1	0	1	1	1	0	0
ɛ	0	1	0	1	1	0	0
œ	0	1	1	1	1	0	0
ɛ̃	0	1	0	1	1	1	0
ɜ	1	1	0	1	1	0	0
ʌ	1	0	0	1	1	0	0
ɔ	1	0	1	1	1	0	0
ɔ̃	1	0	1	1	1	1	0
a	0	1	0	1	0	0	0
Œ	0	1	1	1	0	0	0
ɑ	1	0	0	1	0	0	0
ɒ	1	0	1	1	0	0	0
æ	0	1	0	1	0	0	0
ɐ	1	1	0	1	0	0	0

Appendix C. Model analysis with a continuous response variable

See Tables C.29 and C.30.
See Fig. C.15.

Table C.29
Finnish model continuous accuracy.

	Test items				Training items			
	Est.	SE	$\chi^2(1)$	p	Est.	SE	$\chi^2(1)$	p
(Intercept)	0.28	0.092			0.43	0.010		
ClassSize	0.12	0.092	1.8	0.178	**0.09**	**0.011**	**35.2**	**< 0.001**
TokenFreq	**0.22**	**0.006**	**760.0**	**< 0.001**	**0.13**	**0.002**	**2906.4**	**< 0.001**
Trial	**0.51**	**0.011**	**2218.6**	**< 0.001**	**0.36**	**0.003**	**19477.6**	**< 0.001**
ClassSize:TokenFreq	**−0.01**	**0.004**	**9.1**	**0.003**	**−0.02**	**0.001**	**467.8**	**< 0.001**
TokenFreq:Trial	**−0.31**	**0.011**	**809.5**	**< 0.001**	**−0.23**	**0.003**	**7612.7**	**< 0.001**
ClassSize:Trial	0.02	0.011	2.9	0.09	**−0.08**	**0.003**	**828.5**	**< 0.001**

Note: Linear mixed-effects regression results on 160 test and 1784 training items across ten networks at training trials 60,000 – 350,000.

Table C.30
Polish model continuous accuracy.

	Test items				Training items			
	Est.	SE	$\chi^2(1)$	p	Est.	SE	$\chi^2(1)$	p
(Intercept)	0.232	0.053			0.294	0.010		
ClassSize	0.070	0.052	0.3	0.577	0.005	0.011	0.0	0.971
TokenFreq	**0.284**	**0.004**	**998.4**	**< 0.001**	**0.201**	**0.001**	**8706.4**	**< 0.001**
Trial	**0.464**	**0.004**	**10857.5**	**< 0.001**	**0.478**	**0.001**	**166732.5**	**< 0.001**
ClassSize:TokenFreq	**0.040**	**0.002**	**265.5**	**< 0.001**	**0.006**	**0.001**	**112.8**	**< 0.001**
TokenFreq:Trial	**−0.325**	**0.005**	**5085.0**	**< 0.001**	**−0.231**	**0.001**	**37853.7**	**< 0.001**
ClassSize:Trial	**−0.107**	**0.005**	**549.8**	**< 0.001**	**−0.012**	**0.001**	**99.4**	**< 0.001**

Note: Linear mixed-effects regression results on 192 test and 2168 training items across ten networks at training trials 40,000 – 1,500,000.

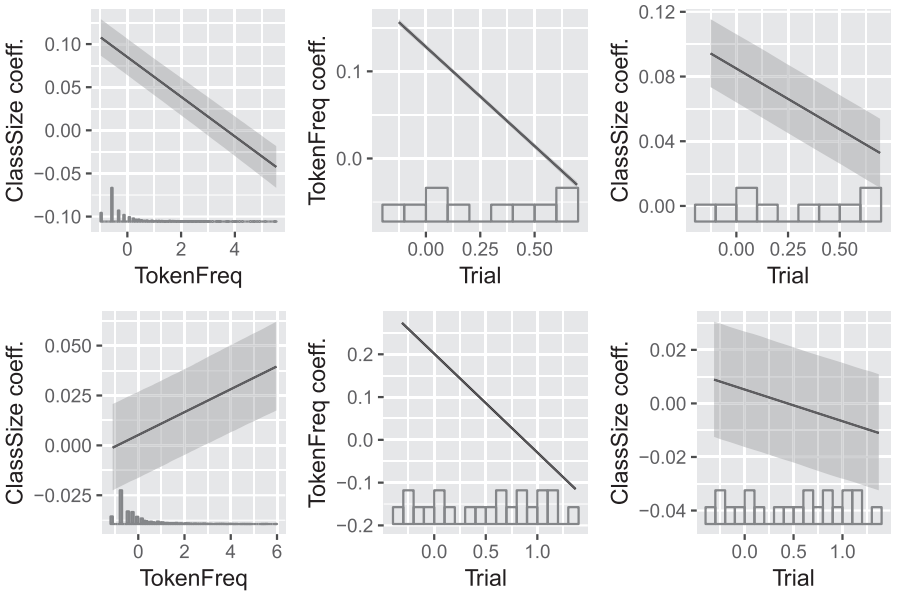

Fig. C.15. Conditional coefficients for two-way interaction terms in linear regression with grammar-based class sizes on Finnish (top row) and Polish (bottom row) model continuous accuracy in training set (Tables C.29 & C.30). Plotted is the *change of the coefficient* of one predictor (y-axis) as a function of the value of the second predictor (x-axis).

Appendix D. Model analysis with corpus-based PND measure

See Tables D.31 and D.32.
See Fig. D.16.

Table D.31
Finnish model accuracy with corpus-based class size measure.

	Test items				Training items			
	Est.	SE	$\chi^2(1)$	p	Est.	SE	$\chi^2(1)$	p
(Intercept)	5.35	0.56			6.55	0.15		
ClassSize	**1.44**	**0.55**	**8.0**	**0.005**	**0.77**	**0.13**	**53.7**	**< 0.001**
TokenFreq	**3.94**	**0.22**	**376.4**	**< 0.001**	**2.22**	**0.09**	**853.8**	**< 0.001**
Trial	**6.18**	**0.73**	**554.0**	**< 0.001**	**7.38**	**0.27**	**2334.6**	**< 0.001**
ClassSize:TokenFreq	−0.15	0.19	0.7	0.418	**−0.31**	**0.06**	**23.6**	**< 0.001**
TokenFreq:Trial	1.82	1.01	3.2	0.072	**1.69**	**0.38**	**20.2**	**< 0.001**
ClassSize:Trial	0.25	0.24	1.2	0.282	**0.23**	**0.11**	**4.3**	**0.038**

Note: Logistic linear mixed-effects regression results on 160 test and 1784 training items across ten networks at training trials 60,000 – 350,000.

Table D.32
Polish model accuracy with corpus-based class size measure.

	Test items				Training items			
	Est.	SE	$\chi^2(1)$	p	Est.	SE	$\chi^2(1)$	p
(Intercept)	3.63	0.37			4.08	0.10		
ClassSize	0.77	0.37	1.6	0.203	**1.07**	**0.10**	**79.7**	**< 0.001**
TokenFreq	**1.61**	**0.07**	**394.2**	**< 0.001**	**1.48**	**0.02**	**3916.6**	**< 0.001**
Trial	**5.02**	**0.15**	**2616.2**	**< 0.001**	**6.23**	**0.06**	**21755.7**	**< 0.001**
ClassSize:TokenFreq	**0.66**	**0.06**	**133.1**	**< 0.001**	**0.19**	**0.01**	**251.7**	**< 0.001**
TokenFreq:Trial	**1.22**	**0.16**	**61.2**	**< 0.001**	**0.81**	**0.07**	**155.3**	**< 0.001**
ClassSize:Trial	**0.69**	**0.07**	**110.4**	**< 0.001**	**0.74**	**0.02**	**929.6**	**< 0.001**

Note: Logistc linear mixed-effects regression results on 192 test and 2431 training items across ten networks at training trials 40,000 – 1,500,000.

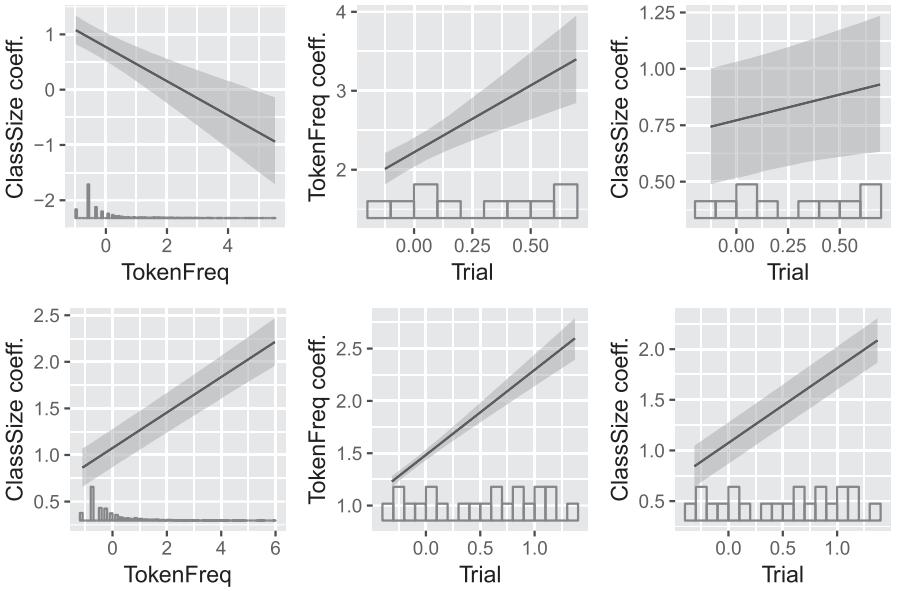

Fig. D.16. Conditional coefficients for two-way interaction terms in linear regression with corpus-based class sizes on Finnish (top row) and Polish (bottom row) model binary accuracy in training set (Tables D.31 & D.32). Plotted is the *change of the coefficient* of one predictor (y-axis) as a function of the value of the second predictor (x-axis).

Appendix E. Details and results of the localist model

Instead of a distributed representation of the stem phonemes, the input layer of the localist models contained 800 units each of which represented a different verb lemma. Together with four extra units encoding person and number, the input layer thus represented each target form in a *pseudo-semantic* fashion. As in the distributed-input models, the output was the inflected phonological form, such that the models' task was a *meaning-to-form* mapping. A disadvantage of this representation is that, unlike children (e.g., Savičiūtė et al., 2018) the model cannot generalise to novel nouns or verbs on the basis of phonological similarity to stored forms. A localist model can, however, still generalise to produce unseen forms of *known* verbs (which is all that is required in the present study), since phonological similarities between — for example — 3sg forms of known verbs are represented on the output layer. Nevertheless, in terms of generalisation, a localist model is at a significant disadvantage because (a) it must fully construct the stem from scratch on the output layer (unlike the distributed model which is given the stem on the input layer) and (b) it must map from a large number of individual input units, rather than a much smaller number of phoneme features.

E.1. Results

Results were in many respects similar to the simulations with the distributed-input models. Three million trials were still sufficient for learning all training forms. Generalisation ability, although stable in Polish (88%), suffered to some extent in Finnish compared to the distributed-input model: generalisation to unknown test forms decreased from 86% to 79%. This is because Finnish has a more complex pattern of stem alternations than Polish, meaning that the Finnish localist model suffers more from the lack of an explicit representation of the stem.

In analogy to the final analysis in Section 4.1, the localist-input models were analysed using a continuous response measure and a corpus-based class size predictor. As shown in Tables E.33 and E.34, both languages showed positive effects of token frequency and training trial, and negative interactions for class size with token frequency as well as class size or token frequency with trial. All of these were predicted and also found in the distributed-input models. In Polish, a main effect of class size was found in both the test set and the training set. There was, however, no significant main effect of class size observed in Finnish, which, again, may be due to a greater difficulty for the model to deal with Finnish stem alternations in the absence of phonological stem information in the input. Class size still had some effect in Finnish in the way of a negative interaction with token frequency in both the test set and the training set.

Table E.33
Finnish localist model continuous accuracy with corpus-based class size measure.

	Test items				Training items			
	Est.	SE	$\chi^2(1)$	p	Est.	SE	$\chi^2(1)$	p
(Intercept)	0.209	0.067			−0.012	0.016		
ClassSize	0.266	0.067	1.6	0.209	−0.006	0.017	0.2	0.661
TokenFreq	**0.409**	**0.013**	**159.5**	**< 0.001**	**0.412**	**0.003**	**333.3**	**< 0.001**
Trial	**0.827**	**0.021**	**1603.8**	**< 0.001**	**1.159**	**0.005**	**51430.7**	**< 0.001**
ClassSize:TokenFreq	**−0.042**	**0.005**	**67.1**	**< 0.001**	**−0.012**	**0.001**	**97.4**	**< 0.001**
TokenFreq:Trial	**−0.603**	**0.021**	**853.4**	**< 0.001**	**−0.694**	**0.005**	**18332.0**	**< 0.001**
ClassSize:Trial	**−0.325**	**0.021**	**247.7**	**< 0.001**	−0.002	0.005	0.1	0.739

Note: Linear mixed-effects regression results on 160 test and 1784 training items across ten networks at training trials 150,000 – 400,000.

Table E.34
Polish localist model continuous accuracy with corpus-based class size measure.

	Test items				Training items			
	Est.	SE	$\chi^2(1)$	p	Est.	SE	$\chi^2(1)$	p
(Intercept)	0.439	0.031			0.243	0.014		
ClassSize	**0.409**	**0.031**	**35.9**	**< 0.001**	**0.095**	**0.015**	**11.6**	**0.001**
TokenFreq	**0.337**	**0.008**	**1000.0**	**< 0.001**	**0.343**	**0.002**	**1519.0**	**< 0.001**
Trial	**0.402**	**0.011**	**1378.4**	**< 0.001**	**0.800**	**0.003**	**91252.6**	**< 0.001**
ClassSize:TokenFreq	**−0.061**	**0.003**	**428.4**	**< 0.001**	**−0.021**	**0.001**	**1242.4**	**< 0.001**
TokenFreq:Trial	**−0.368**	**0.011**	**1139.7**	**< 0.001**	**−0.458**	**0.003**	**29250.9**	**< 0.001**
ClassSize:Trial	**−0.324**	**0.011**	**885.8**	**< 0.001**	**−0.064**	**0.003**	**568.8**	**< 0.001**

Note: Linear mixed-effects regression results on 192 test and 2431 training items across ten networks at training trials 150,000 – 700,000.

Frequency-based substitutions (shown in Tables E.35 and E.36) occurred less often with localist input. In Finnish, substitutions occurred only in 2pl contexts (where passives were produced instead); in Polish only in 1sg contexts (where 3sg forms were produced instead). Conjugation class errors (Tables E.37 and E.38) were similar to the distributed-input models.

Table E.35
Finnish localist model: Person/number substitution errors in 160 test items after 1000 training trials (suffix mean accuracy at 75%).

	Output	
Target	1sg	pass
2pl	2	313

Table E.36
Polish localist model: Person/number substitution errors in 192 test items after 20,000 training trials (suffix mean accuracy at 75%).

	Output
Target	3sg
1sg	14

Table E.37
Finnish localist model: Conjugation class (overgeneralisation) errors in 160 test items after 1000 training trials (suffix mean accuracy at 75%). Only unambiguous errors counted, i.e., where the output ending was assignable to exactly one class.

	Output	
Target	II	III
I	9	0
II	0	1
III	8	0
IV	8	0

Table E.38
Polish localist model: Conjugation class (overgeneralisation) errors in 160 test items after 20,000 training trials (suffix mean accuracy at 75%). Only unambiguous errors counted, i.e., where the output ending was assignable to exactly one class.

	Output		
	II	III	IV
I	47	27	1
III	6	0	1
IV	1	12	0

See Fig. E.17.

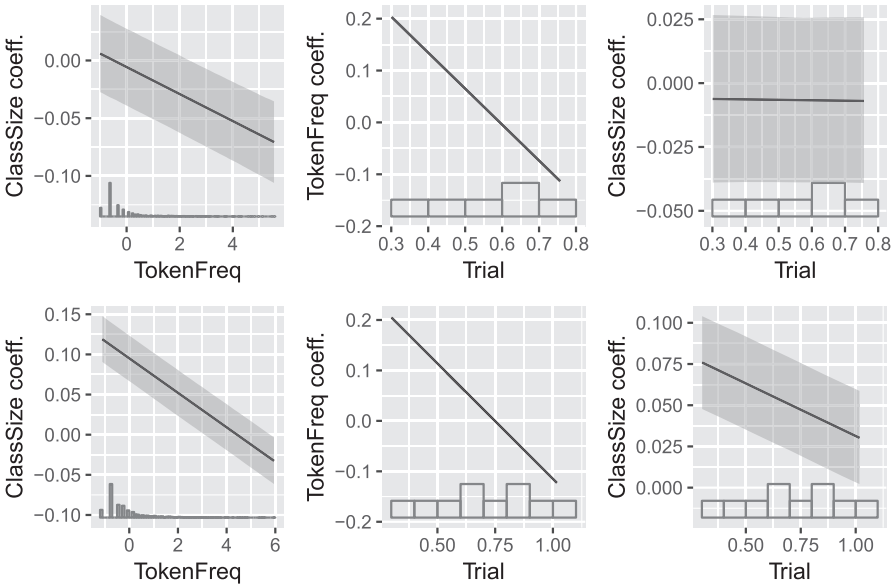

Fig. E.17. Conditional coefficients for two-way interaction terms in linear regression with corpus-based class sizes on Finnish (top row) and Polish (bottom row) model continuous accuracy in training set (Tables E.33 & E.34). Plotted is the *change of the coefficient* of one predictor (y-axis) as a function of the value of the second predictor (x-axis).

Appendix F. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.cogpsych.2019.02.001.

References

Aguado-Orea, J. J. (2004). *The acquisition of morpho-syntax in Spanish: Implications for current theories of development.* PhD thesisUniversity of Nottingham United Kingdom.

Aguado-Orea, J., & Pine, J. M. (2015). Comparing different models of the development of verb inflection in early child spanish. *PLoS ONE, 10*(3).

Albright, A., & Hayes, B. (2003). Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition, 90*(2), 119–161.

Alegre, M., & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language, 40*(1), 41–61.

Ambridge, B. (2010). Children's judgments of regular and irregular novel past-tense forms: New data on the english past-tense debate. *Developmental Psychology, 46*(6), 1497.

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language, 42*(2), 239–273.

Andrews, S. (1982). Phonological recoding: Is the regularity effect consistent? *Memory & Cognition, 10*(6), 565–575.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412.

Barner, D., Wood, J., Hauser, M., & Carey, S. (2008). Evidence for a non-linguistic distinction between singular and plural sets in rhesus monkeys. *Cognition, 107*(2), 603–622.

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Berko, J. (1958). The child's learning of English morphology. *Word, 14*(2-3), 150–177.

Bertram, R., Laine, M., & Virkkala, M. M. (2000). The role of derivational morphology in vocabulary acquisition: Get by with a little help from my morpheme friends. *Scandinavian Journal of Psychology, 41*(4), 287–296.

Bittner, D., Dressler, W. U., & Kilani-Schoch, M. (2003). *Development of verb inflection in first language acquisition: A cross-linguistic perspective, vol. 21*. Walter de Gruyter.

Blything, R. P., Ambridge, B., & Lieven, E. V. (2018). Children's acquisition of the English past-tense: Evidence for a single-route account from novel verb production data. *Cognitive Science*.

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology, 23*(1), 45–93.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form, vol. 9*. John Benjamins Publishing.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes, 10*(5), 425–455.

Bybee, J. L., & Moder, C. L. (1983). Morphological classes as natural categories. *Language,* 251–270.

Chandler, S. (2010). The english past tense: Analogy redux. *Cognitive Linguistics, 21*(3), 371–417.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York: Harper & Row.

Clahsen, H., Rothweiler, M., Woest, A., & Marcus, G. F. (1992). Regular and irregular inflection in the acquisition of german noun plurals. *Cognition, 45*(3), 225–255.

Cottrell, G. W., & Plunkett, K. (1994). Acquiring the mapping from meaning to sounds. *Connection Science, 6*(4), 379–412.

Daugherty, K., & Seidenberg, M. (1992). Rules or connections? The past tense revisited. *Annual conference of the cognitive science society: vol. 14*, (pp. 259–264).

Daugherty, K. G., & Seidenberg, M. S. (1994). Beyond rules and exceptions. *The Reality of Linguistic Rules, 26*, 353.
Deen, K. U. (2004). Productive agreement in Swahili: Against a piecemeal approach. In A. Brugos, M. R. Clark-Cotton, & S. Ha (Eds.). *Proceedings of the 29th annual Boston University conference on language development* (pp. 156–167). Boston: Cascadilla Press.
Deen, K. U., & Hyams, N. (2006). The morphosyntax of mood in early grammar with special reference to swahili. *First Language, 26*(1), 67–102.
Dąbrowska, E. (2004). Rules or schemas? Evidence from Polish. *Language and Cognitive Processes, 19*(2), 225–271.
Dąbrowska, E. (2005). Productivity and beyond: Mastering the polish genitive inflection. *Journal of Child Language, 32*(1), 191–205.
Dąbrowska, E. (2008a). The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language, 58*(4).
Dąbrowska, E. (2008b). The later development of an early-emerging system: The curious case of the Polish genitive. *Linguistics, 46*(3), 629–650.
Dąbrowska, E., & Szczerbiński, M. (2006). Polish children's productivity with case marking: The role of regularity, type frequency, and phonological diversity. *Journal of Child Language, 33*(3), 559–597.
Dąbrowska, E., & Tomasello, M. (2008). Rapid learning of an abstract language-specific category: Polish children's acquisition of the instrumental construction. *Journal of Child Language, 35*(3), 533–558.
Drabik, L. & Sobol, E. (2007). *Słownik Języka Polskiego PWN*. Warszawa: Wydawnictwo Naukowe PWN.
Dryer, M.S. & Haspelmath, M. (Eds.) (2013). *The World Atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Availale at <http://wals.info/>.
Elman, J. L. (1998). *Rethinking innateness: A connectionist perspective on development, vol. 10*. MIT Press.
Forrester, N., & Plunkett, K. (1994). The inflectional morphology of the Arabic broken plural: A connectionist account. *Proceedings of the sixteenth annual meeting of the cognitive science society*. Hillsdale, NJ: Erlbaum.
Gabry, J. & Goodrich, B. (2016). *rstanarm: Bayesian applied regression modeling via Stan*. R package version 2.10.1.
Gallup, G. G., Jr. (1998). Self-awareness and the evolution of social intelligence. *Behavioural Processes, 42*(2–3), 239–247.
Glanzer, M. (1982). Short-term memory. *Handbook of research methods in human memory and cognition* (pp. 63–98). Elsevier.
Granlund, S., Kolak, J., Vihman, V., Engelmann, F., Ambridge, B., Pine, J.,..., & Lieven, E. (2018). Language-general and language-specific phenomena in the acquisition of inflectional noun morphology: A cross-linguistic elicited-production study of Polish, Finnish and Estonian. submitted for publication.
Gvozdev, A. (1949). Formirovaniye u rebenka grammaticheskogo stroya russkogo yazyka. *Moscow: Izd-vo Akademia Pedagogicheskikh Nauk RSFSR*.
Hahn, U., & Nakisa, R. C. (2000). German inflection: Single route or dual route? *Cognitive Psychology, 41*(4), 313–360.
Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T., & Alho, I. (2004). *Iso suomen kielioppi*. Helsinki: Suomen Kirjallisuuden Seura.
Haman, E., Etenkowski, B., Łuniewska, M., Szwabe, J., Dąbrowska, E., Szreder, M., & Łaziński, M. (2011). *The Polish CDS corpus*.
Hare, M., & Elman, J. L. (1995). Learning and morphological change. *Cognition, 56*(1), 61–98.
Hare, M., Elman, J. L., & Daugherty, K. G. (1995). Default generalisation in connectionist networks. *Language and Cognitive Processes, 10*(6), 601–630.
Harris, T., & Wexler, K. (1996). The optional infinitive stage in child English. In H. Clahsen (Ed.). *Generative perspectives on language acquisition* (pp. 1–42). Amsterdam: John Benjamins.
Hartshorne, J. K., & Ullman, M. T. (2006). Why girls say 'holded' more than boys. *Developmental Science, 9*(1), 21–32.
Hay, J. (2018). Sociophonetics: The role of words, the role of context, and the role of words in context. *Topics in Cognitive Science*.
Hoekstra, T., & Hyams, N. (1998). Aspects of root infinitives. *Lingua, 106*(1–4), 81–112.
Joanisse, M. F., & McClelland, J. L. (2015). Connectionist perspectives on language learning, representation and processing. Wiley interdisciplinary reviews. *Cognitive Science, 6*(3), 235–247.
Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences, 96*(13), 7592–7597.
Juliano, C., & Tanenhaus, M. K. (1994). A constraint-based lexicalist account of the subject/object attachment preference. *Journal of Psycholinguistic Research, 23*(6), 459–471.
Kenstowicz, M., & Kisseberth, C. (2014). *Generative phonology: Description and theory*. Academic Press.
Kirjavainen, M., Kidd, E., & Lieven, E. (2017). How do language-specific characteristics affect the acquisition of different relative clause types? Evidence from Finnish. *Journal of Child Language, 44*(1), 120–157.
Kirjavainen, M., Nikolaev, A., & Kidd, E. (2012). The effect of frequency and phonological neighbourhood density on the acquisition of past tense verbs by Finnish children. *Cognitive Linguistics, 23*(2), 273–315.
Krajewski, G., Lieven, E. V., & Theakston, A. L. (2012). Productivity of a Polish child's inflectional noun morphology: A naturalistic study. *Morphology, 22*(1), 9–34.
Kunnari, S., Savinainen-Makkonen, T., Leonard, L. B., Mäkinen, L., Tolonen, A.-K., Luotonen, M., & Leinonen, E. (2011). Children with specific language impairment in Finnish: The use of tense and agreement inflections. *Journal of Child Language, 38*(5), 999–1027.
Legate, J. A., & Yang, C. (2007). Morphosyntactic learning and the development of tense. *Language Acquisition, 14*(3), 315–344.
Leonard, L. B., Caselli, M. C., & Devescovi, A. (2002). Italian children's use of verb and noun morphology during the preschool years. *First Language, 22*(3), 287–304.
Leonard, L. B., Kas, B., & Pléh, C. (2009). The use of tense and agreement by Hungarian-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research, 52*(1), 98–117.
Li, P., & MacWhinney, B. (2002). PatPho: A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments, & Computers, 34*(3), 408–415.
MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review, 109*(1), 35–54.
MacWhinney, B. (1978). The acquisition of morphophonology. *Monographs of the Society for Research in Child Development*, 1–123.
MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition, 40*(1), 121–157.
Marchman, V. A. (1997). Children's productivity in the English past tense: The role of frequency, phonology, and neighborhood structure. *Cognitive Science, 21*(3), 283–304.
Marchman, V. A., Wulfeck, B., & Weismer, S. E. (1999). Morphological productivity in children with normal language and SLI: A study of the English past tense. *Journal of Speech, Language, and Hearing Research, 42*(1), 206–219.
Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. MIT Press.
Maslen, R. J., Theakston, A. L., Lieven, E. V., & Tomasello, M. (2004). A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language, and Hearing Research, 47*(6), 1319–1333.
Matthews, D. E., & Theakston, A. L. (2006). Errors of omission in English-speaking children's production of plurals and the past tense: The effects of frequency, phonology, and competition. *Cognitive Science, 30*(6), 1027–1052.
McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: what does the evidence rule out? *Trends in Cognitive Sciences, 6*(11), 465–472.
Mielikäinen, A. (1984). Monikon 3.persoonan kongruenssi puhekielessä [Concord of 3rd person plural in present-day spoken Finnish]. *Virittäjä, 13*, 162–175.
Mirković, J., Seidenberg, M. S., & Joanisse, M. F. (2011). Rules versus statistics: Insights from a highly inflected language. *Cognitive Science, 35*(4), 638–681.
Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23*(1), 103–123.
Murdock, B. B., Jr (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*(5), 482.
Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas—part II. *Language and Linguistics Compass, 10*(11), 591–613.
O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.
Oxford English Dictionary (2018). *How many words are there in the english language?* Retrieved from <https://en.oxforddictionaries.com/explore/how-many-words-are-there-in-the-english-language> (19th January, 2018).
Pearlmutter, N. J., & MacDonald, M. C. (1995). Individual differences and probabilistic constraints in syntactic ambiguity resolution. *Journal of Memory and Language, 34*(4), 521.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. *Typological Studies in Language, 45*, 137–158.
Pinker, S. (1984). *Language learnability and language development, vol. 1*. Harvard University Press.
Pinker, S. (1998). Words and rules. *Lingua, 106*(1–4), 219–242.
Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*(1), 73–193.
Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences, 6*(11), 456–463.
Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*(1), 56–115.
Plunkett, K., & Bandelow, S. (2006). Stochastic approaches to understanding dissociations in inflectional morphology. *Brain and Language, 98*(2), 194–209.
Plunkett, K., & Juola, P. (1999). A connectionist model of english past tense and plural morphology. *Cognitive Science, 23*(4), 463–490.
Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition, 38*(1), 43–102.
Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition, 48*(1), 21–69.
Plunkett, K., & Marchman, V. A. (1996). Learning from a connectionist model of the acquisition of the English past tense. *Cognition, 61*(3), 299–308.
Plunkett, K., & Nakisa, R. C. (1997). A connectionist model of the Arabic plural system. *Language and Cognitive Processes, 12*(5-6), 807–836.
Plunkett, K., Sinha, C., Møller, M. F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science, 4*(3-4), 293–312.
Pothos, E. M. (2005). The rules versus similarity distinction. *Behavioral and Brain Sciences, 28*(1), 1–14.
Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes, 8*(1), 1–56.
Räsänen, S. H. M., Ambridge, B., & Pine, J. M. (2016). An elicited-production study of inflectional verb morphology in child Finnish. *Cognitive Science, 40*, 1704–1738.
R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Rogers, T. T., & McClelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cognitive Science, 38*(6), 1024–1077.
Rubino, R. B., & Pine, J. M. (1998). Subject–verb agreement in brazilian portuguese: What low error rates hide. *Journal of Child Language, 25*(1), 35–59.
Ruh, N., & Westermann, G. (2008). A single-mechanism dual-route model of German verb inflection. *Proceedings of the 30th annual conference of the cognitive science society* (pp. 2209–2216). .
Rumelhart, D.E. and McClelland, J.L. (1985). *On learning the past tenses of English verbs*. Technical report, DTIC Document.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive Modeling, 323*, 533–536.
Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Volume 1. Foundations*. Cambridge, MA: MIT Press.
Saloni, Z. (1976). *Cechy składniowe polskiego czasownika, vol. 76*. Ossolińskich: Zakład Narodowy im.
Savičiūtė, E., Ambridge, B., & Pine, J. M. (2018). The roles of word-form frequency and phonological neighbourhood density in the acquisition of Lithuanian noun morphology. *Journal of Child Language, 45*(3), 641–672.
Schuler, K. D., Yang, C., & Newport, E. L. (2016). Testing the tolerance principle: Children form productive rules when it is more computationally efficient to do so. *CogSci*.
Seidenberg, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition, 19*(1), 1–30.
Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*(4), 523.
Seidenberg, M. S., Waters, G. S., Barnes, M. A., & Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Memory and Language, 23*(3), 383.
Slobin, D. I. (1968). Early grammatical development in several languages, with special attention to Soviet research. *Working Papers*, 11.
Slobin, D. I. (1973). Cognitive prerequisites for the development of grammar. *Studies of Child Language Development, 1*, 75–208.
Slobin, D. I., & Bever, T. G. (1982). Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition, 12*(3), 229–265.
Smoczyńska, M. (1985). The acquisition of Polish. In D. I. Slobin (Ed.). *The crosslinguistic study of language acquisition* (pp. 595–686). Hillsdale, N.J: Erlbaum.
Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology, 12*(3), 175–200.
Stan Development Team (2015). *Stan: A C++ Library for Probability and Sampling, Version 2.10.0*.
Stavrakaki, S., & Clahsen, H. (2009). The perfective past tense in greek child language. *Journal of Child Language, 36*(1), 113–142.
Stephany, U., & Voeikova, M. D. (2009). *Development of nominal inflection in first language acquisition: A cross-linguistic perspective, vol. 30*. Walter de Gruyter.
Stoll, S., Bickel, B., Lieven, E., Paudyal, N. P., Banjade, G., Bhatta, T. N., ... Rai, M. (2012). Nouns and verbs in Chintang: children's usage and surrounding adult speech. *Journal of Child Language, 39*(2), 284–321.
Suomi, K., Toivanen, J., & Ylitalo, R. (2008). Finnish sound structure. *Studia Humaniora Ouluensia, 9*.
Swan, O. E. (2003). *Polish grammar in a nutshell*. Unpublished material.
Taraban, R., & McClelland, J. L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and language, 26*(6), 608.
Tatsumi, T., Ambridge, B., & Pine, J. M. (2017). Disentangling effects of input frequency and morphophonological complexity on children's acquisition of verb inflection: An elicited production study of Japanese. *Cognitive Science*.
Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2005). The acquisition of auxiliary syntax: Be and have. *Cognitive Linguistics, 16*(1), 247–277.
Theakston, A. L., Lieven, E. V., & Tomasello, M. (2003). The role of the input in the acquisition of third person singular verbs in English. *Journal of Speech, Language, and Hearing Research, 46*(4), 863–877.
Theakston, A. L., & Rowland, C. F. (2009). The acquisition of auxiliary syntax: A longitudinal elicitation study. Part 1: Auxiliary be. *Journal of Speech, Language, and Hearing Research, 52*(6), 1449–1470.
Tokarski, J. (1951). *Czasowniki polskie: Formy, typy, wyjątki, słownik*. Wydawn. S. Arcta.
Tomasello, M. (2009). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
Waters, G. S., & Seidenberg, M. S. (1985). Spelling-sound effects in reading: Time-course and decision criteria. *Memory & Cognition, 13*(6), 557–572.
Westermann, G., & Ruh, N. (2012). A neuroconstructivist model of past tense development and processing. *Psychological Review, 119*(3), 649–667.
Wexler, K. (1998). Very early parameter setting and the unique checking constraint: A new explanation of the optional infinitive stage. *Lingua, 106*(1–4), 23–79.
Yang, C. D. (2002). *Knowledge and learning in natural language*. Oxford University Press on Demand.